G Clustering Is Important for the Initiation of Transcription-Induced R-Loops In Vitro, whereas High G Density without Clustering Is Sufficient Thereafter\textsuperscript{V}\textsuperscript{†}

Deepankar Roy and Michael R. Lieber*

Departments of Pathology, Biochemistry and Molecular Biology, and Molecular Microbiology and Immunology, Biological Sciences, USC Norris Comprehensive Cancer Ctr., Rm. 5428, 1441 Eastlake Ave., MC9176, Los Angeles, California 90089-9176

Received 30 January 2009/Returned for modification 27 February 2009/Accepted 13 March 2009

R-loops form cotranscriptionally in vitro and in vivo at transcribed duplex DNA regions when the nascent RNA is G-rich, particularly with G clusters. This is the case for phage polymerases, as used here (T7 RNA polymerase), as well as RNA polymerases in bacteria, Saccharomyces cerevisiae, avians, mice, and humans. The nontemplate strand is left in a single-stranded configuration within the R-loop region. These structures are known to form at mammalian immunoglobulin class switch regions, thus exposing regions of single-stranded DNA for the action of AID, a single-strand-specific cytidine deaminase. R-loops form by thread-back of the RNA onto the template DNA strand, and here we report that G clusters are extremely important for the initiation phase of R-loop formation. Even very short regions with one GGGG sequence can initiate R-loops much more efficiently than random sequences. The high efficiencies observed with G clusters cannot be achieved by having a very high G density alone. Annealing of the transcript, which is otherwise disadvantaged relative to the nontemplate DNA strand because of unfavorable proximity while exiting the RNA polymerase, can offer greater stability if it occurs at the G clusters, thereby initiating an R-loop. R-loop elongation beyond the initiation zone occurs in a manner that is not as reliant on G clusters as it is on a high G density. These results lead to a model in which G clusters are important to nucleate the thread-back of RNA for R-loop initiation and, once initiated, the elongation of R-loops is primarily determined by the density of G on the nontemplate DNA strand. Without both a favorable R-loop initiation zone and elongation zone, R-loop formation is inefficient.

Immunoglobulin (Ig) class switch recombination (CSR) is the process in which IgM is changed to IgG, IgA, or IgE by rearranging the Ig heavy chain from IgH\textsubscript{a} to IgH\textsubscript{γ}, IgH\textsubscript{δ}, or IgH\textsubscript{ε} (7, 11, 40). This DNA recombination process occurs at class switch sequences located upstream of the corresponding constant domain exons. The class switch sequences are long (1 to 12 kb); repetitive, with unit repeat lengths of 25 to 80 bp; transcribed by a promoter immediately upstream of each switch region; and strikingly G-rich on the nontemplate strand, with G densities reaching 40 to 50\% (48). Despite conservation of these features, the actual primary switch repeat sequences themselves are not conserved across species or even among the switch sequences of the different isotypes (e.g., Ig\textsubscript{μ}, Ig\textsubscript{γ}, Ig\textsubscript{κ}, and Ig\textsubscript{λ}) (15). Even the unit repeats within any one switch region of a given species vary from one repeat to the next, such that not one of the individual repeats precisely matches the average sequence of that switch region.

Activation-induced deaminase (AID) is a cytidine deaminase that is expressed in activated B cells and is essential for Ig CSR and Ig somatic hypermutation (SHM) (31). AID only acts on cytosines located in single-stranded DNA (ssDNA) (6). This raises the question of how the DNA becomes single-stranded so that AID can act on these genomic regions (48). The promoters upstream of the switch region are critical for CSR, indicating that transcription is critical (5, 16, 21, 46, 49). Transcription is also critically important for SHM, suggesting that some level of ssDNA is somehow exposed to AID during transcription (28). Indeed, transcription by both mammalian RNA polymerase II and prokaryotic or phage polymerases can generate ssDNA upstream of the polymerase to some extent in a manner that is not very well characterized (29). The eukaryotic ssDNA binding protein, RPA, appears to contribute to this exposure of ssDNA in vivo, perhaps by stabilizing the single-stranded state transiently induced by transcription (7, 8, 32, 33). Other proteins that bind either the nascent RNA or the nontemplate DNA strand may modify the efficiency of R-loop formation in vivo (14, 20, 22, 26).

Ig switch regions evolved several hundred million years after SHM already existed (4, 19, 40, 42, 50). Although both SHM and CSR require AID, the processes of CSR and SHM are quite different. SHM is a point mutagenesis process, whereas CSR is a double-strand break recombination process. CSR regions in Xenopus are rich in palindromic forms of preferred sites of AID action (WRC, where W = A or T and R = A or G). Interestingly, upon hyperimmunization, amphibians do not switch nearly as efficiently as mammals (13). At least part of the basis of this may be due to the high asymmetric G density in mammals. Mammalian CSR regions achieve G densities of nearly 50\% on the nontemplate strand in some repeats, in contrast to amphibian switch regions, which are 21\% G, like random vertebrate DNA.

\* Corresponding author. Mailing address: USC Norris Comprehensive Cancer Ctr., Rm. 5428, 1441 Eastlake Ave., MC9176, Los Angeles, CA 90089-9176. Phone: (323) 865-0568. Fax: (323) 865-3019. E-mail: lieber@usc.edu.

† Supplemental material for this article may be found at http://mcb.asm.org/.

\textsuperscript{V} Published ahead of print on 23 March 2009.
An RNA-DNA hybrid forms at a 140-bp subregion of the Igα switch region upon in vitro transcription with T7 RNA polymerase (34, 35). We showed that RNA-DNA hybrids form at all of the tested murine switch regions anywhere within the length of their repetitive regions (9). The RNA-DNA hybrids are stable for days and stable to phenol-chloroform extraction. The RNA-DNA hybrids are also stable to shorter term exposure to temperatures of 65 to 75°C. We showed that the structure of these RNA-DNA hybrids is an R-loop, with the G-rich DNA strand displaced by the G-rich nascent RNA of the same sequence (47). We and others have shown that the number of hydrogen bonds between the RNA and template DNA strand is important for the stability of the R-loop, based on failure of R-loops to form when inosine is substituted for guanine (1:C base pairs share only two hydrogen bonds rather than the three of G·C) (12, T. E. Wilson and M. R. Lieber, unpublished data). Upon RNase H treatment, some misalignment of top strand and bottom strand DNA repeats occurs in vitro, resulting in displaced loops of ssDNA on both strands, and we have proposed that this may occur in vivo as a way to expose single-strandedness on the template DNA strand for AID action (47, 48).

Our in vitro studies do not show any evidence of secondary structure on the nontemplate DNA strand of the R-loops (e.g., G-quartets), and in vitro experiments in which no Na+ or K+ are present (only lithium ion or cesium ion) show unaltered levels of R-loop formation, indicating that R-loops are stable under conditions where G-quartets do not form (37). Therefore, if G-quartets form in vitro on the G-rich nontemplate DNA strand, they are not essential for R-loop formation. Moreover, at chromosomal R-loops, the regions of single-strandedness are continuous (47, 48). One would expect G-quartets to trap intervening Cs, thereby making them resistant to bisulfite, and we do not see this (17). For these in vitro and in vivo reasons, we do not favor the view that G-quartets exist at switch region R-loops; however, one cannot rule out the possibility that G-quartets exist but that physical methods to detect them are limited (12).

We and others have demonstrated that R-loops form at Ig CSR regions in the mouse chromosome in activated B cells as well as in vitro and in vivo (12, 26). Ethidium bromide staining was done afterward to locate restriction fragments that were digested. In vitro transcription, sodium bisulfite treatment, and RNase H digestion were performed. To calculate the frequencies of R-loop formation, samples of T7 transcribed substrates were electrophoresed and transferred to nylon membranes. The oligonucleotide insert was hybridized to the membranes and autoradiography was used to determine the number of restriction fragments that were not digested.

In vitro transcription. In vitro transcription, sodium bisulfite treatment, and frequency determination experiments using a colony lift hybridization assay were performed as previously described (37). Briefly, Sall-linearized substrates were mock transcribed or transcribed with T7 RNA polymerase in the presence of [α-32P]UTP for 1 h at 37°C. Transcribed samples were treated with RNase A or RNase A and RNase H1 for one additional hour at 37°C, organically extracted, and electrophoresed. Ethidium bromide staining was done afterward to locate restriction fragments containing the switch regions and the shifted bands. The gels were exposed to phosphorimager screens after pressing and drying as described previously.

Colony lift hybridization for the determination of frequency of R-loop formation. To calculate the frequencies of R-loop formation, samples of T7 transcribed and RNase A-treated substrates were incubated overnight with sodium bisulfite
at 37°C without any denaturation step. This would convert Cs on the single-stranded regions on nontemplate and template DNA in the R-loop conformation. PCR amplification was done on bisulfite-modified DNA with native primers DR050 and DR051 that would anneal outside of the region of interest (37). The PCR fragments were cloned with TOPO-TA cloning kit, and the white recombinant bacterial colonies were restreaked and lifted onto nylon membranes. The membranes were then transferred to a denaturing solution (0.5 M NaOH, 1.5 M NaCl) for 15 min and then transferred to 1 M Tris (pH 7.5) for 15 min. This was followed by transfer to 1 M Tris (pH 7.5)–1.5 M NaCl for 15 min and a rinse with 2× SSC (1× SSC is 0.15 M NaCl plus 0.015 M sodium citrate). The DNA was then fixed on the membrane by UV cross-linking. The membranes were then rinsed with 2× SSC and incubated for hybridization with end-labeled oligonucleotide probes. The oligonucleotide probes were designed to anneal to regions of C-to-T conversion but not to unconverted regions in the nontemplate strand. The oligonucleotide probes were designed to anneal to regions of C-to-T conversions but not to unconverted regions in the nontemplate strand. The oligonucleotide probes were designed to anneal to regions of C-to-T conversions but not to unconverted regions in the nontemplate strand. The oligonucleotide probes were designed to anneal to regions of C-to-T conversions but not to unconverted regions in the nontemplate strand.

RESULTS

G Clustering is important for R-loop initiation. In our previous work, we determined that the density of Gs is not the only parameter that is important for R-loop formation because clustering of Gs yielded more efficient R-looping (37). For example, we noted that for the same G density, clustered Gs supported more R-loop formation than dispersed G-rich regions (7% of molecules with R-loops versus 0.3% [see Fig. 6A in reference 37]). Given this, we wondered whether the addition of one or two additional GGGG clusters at the very beginning of an RNA transcript might increase the percentage of DNA templates that form R-loops. Therefore, we varied the number and location of Gs in the sequence in the manner diagrammed in Fig. 1. Nearly all of the sequence shown in the top line (substrate pDR18A) consists of four ~49-bp repeats of the Sy3 switch region, but with the addition of a 16-bp sequence before these four repeats. We use three variations of this 16-bp sequence, called A, C, and B. The A variant of the 16-bp sequence contains two GGGG stretches on the nontemplate strand. The C variant has only one cluster of GGGG, and the B variant has none. We examined how these variants influence R-loop formation upon transcription from the T7 promoter with extension through the 4 repeats immediately downstream.

For the experiments, we incubated T7 RNA polymerase with purified, linearized plasmid DNA for 1 h at 37°C (see Fig. 1 in reference 37). RNase A is added to degrade free RNA after transcription. The reactions are then organically extracted and analyzed on agarose gels. The RNA generated remains associated with the template DNA in the form of an R-loop in a subset of molecules when the RNA strand is sufficiently G-rich. For these linear templates, the fraction of molecules in an R-loop conformation can be seen as a shifted species due to its slower gel mobility (9, 10, 30, 34, 47, 48).

Based on the ethidium-stained gels (Fig. 2A), we found that upon T7 transcription using pDR18A, pDR18C, and pDR18B, the pDR18A variant exhibits the greatest amount of shift (i.e., R-loop formation), whereas the shift using pDR18C (with one GGGG cluster, abbreviated 1x4G) is comparatively weaker. The amount of shift drops even further for the pDR18B variant, which contains no G clusters. The transcription is done with [α-32P]UTP present. Therefore, after RNase A treatment, the shifted species can be visualized by phosphorimaging (Fig. 2B), whereas the linearized fragment containing the transcribed sequence, but not associated with RNA, remains largely unlabeled. The densitometric analysis of the radiolabeled shifted species also reveals a similar trend in that the radiolabeled shift in the substrate with the A motif (2x4G) is severalfold greater than the shift in the substrate with the B motif (random sequence; no G clusters, or 0x4G). The substrate with the C motif (1x4G) also shows an increase over the B variant. RNase H1 digests the RNA in RNA-DNA hybrids, and treatment of a fraction of transcribed sample with E. coli RNase H1 confirmed the RNA-DNA hybrid nature of the...
FIG. 2. Effect of G clusters in the RIZ on R-loop formation. Analysis and map of R-loop molecules in pDR18A, pDR18C, and pDR18B with RIZ motif A (2 × GGGG), C (1 × GGGG), or B (no G clusters) are shown. These plasmids have identical REZ regions. (A) Linearized pDR18A, pDR18C, and pDR18B substrates were either mock transcribed (lanes 1, 6, and 11 for pDR18A, pDR18C, and pDR18B, respectively) or transcribed with T7 RNA polymerase in the presence of [α-32P]UTP and treated with RNase A afterward (lanes 2 to 4, lanes 7 to 9, and lanes 12 to 14 in triplicate for pDR18A, pDR18C, and pDR18B, respectively). The fifth lane in each set is a transcribed sample treated with RNase A and RNase H1 (lanes 5, 10, and 15 in Fig. 2A and 2B).

To calculate the frequency of R-looped molecules and map the sequence location of the R-loops generated in these substrates, we treated the transcribed samples with sodium bisulfite (as described in Materials and Methods) followed by a colony lift hybridization assay (see Materials and Methods) to screen and sequence many molecules for R-loops. We observed the same trend in R-loop frequencies: 12% for the A variant (2x4G), 6% for the C variant (1x4G), and 0.8% for the B variant (0x4G), as measured by colony lift hybridization. We find that nearly all of the R-loops of the A variants begin within a narrow region 10 to 26 nt downstream of the transcription initiation site within the G clusters of the RIZ (Fig. 2C).

The sequence downstream of the RIZ is exactly the same in the three variants (A, C, and B); therefore, the difference in R-loop frequencies is a direct consequence of the composition of the RIZ. Having two G clusters in the RIZ imparts better efficiency than having one, which in turn is better than none. The thermodynamic stability of an RNA-DNA hybrid that initiates at the start of the RIZ and extends to the end of the four-switch region repeats is affected by <1.05-fold in the A or C variants compared to the B variant (see Tables S2 to S4, column 6, in the supplemental material), and yet we see R-looping variation over a 15-fold range between these three substrates. Therefore, the difference in R-loop formation efficiency is due to the differences only in the RIZ rather than the sequences downstream.

G clustering in the initiation zone is important for a wide range of substrates. We wanted to extend analysis of the RIZ effect on R-loop formation to a range of substrates where the sequence downstream, termed the R-loop elongation zone or REZ, has been mutated. In these mutated forms, all clusters of 4 Gs are reduced to either clusters of 3 Gs (pDR22A, pDR22C, and pDR22B), clusters of 2 Gs (pDR26A, pDR26C, and pDR26B), or only dispersed Gs with no G clusters but relatively high G density (pDR54A, pDR54C, and pDR54B). The RIZ region is one of the 3 sequences described above (A, C, and B). The substrates thus constructed were linearized, "Shift" bands, but not with the "L" bands, and is not seen in the mock-transcribed lanes or in the RNase H-treated samples at either position. (C) Representation of single-stranded regions in the DNA nontemplate strand. Transcribed substrates were treated with sodium bisulfite to convert Cs in the single-stranded regions to Us. PCR amplification, cloning, and colony lift hybridization were done to calculate R-loop frequency (also see Table 1) and detect regions of single-strandedness (read as stretches of C-to-T conversions with sequencing). The top line is a diagram of the linearized substrate, showing the T7 promoter, followed by the RIZ sequence A, C, or B upstream (shown as an inverted triangle) of the REZ switch repeats represented as thick arrows. In each set, the first line shows all Cs on the nontemplate strand as vertical lines. Each of the following lines represents an independent nontemplate strand derivative molecule, with vertical lines representing observed as C-to-T conversions. Some molecules with R-loop-induced single-stranded stretches of conversion were incomplete for the conversion information on the nontemplate strand, and only the length to which the molecule was informative for the nontemplate strand has been shown. The asterisks mark the position of the internal C in the CCGGmetATGG sequence that gets methylated by bacterial dcm (DNA cytosine methylase) enzyme and therefore remains unconverted upon sodium bisulfite treatment.
transcribed with T7 RNA polymerase and treated with RNase A, or treated with RNase A and RNase H1.

In the substrates with only 3G clusters in the REZ (pDRA, pDRC, and pDRB), we observed that the A variant exhibited a marked improvement in R-loop formation, as seen on the ethidium-stained gel, on the radiolabeled shift, and in the colony lift assay. In the ethidium-stained gel, the shifted species was ca. 5.3% of the total transcribed substrate (Fig. 3A, lanes 2 to 4). Although a distinct shifted species could not be detected for the C or the B variants in the ethidium-stained gel, in a more sensitive assay using the [α-32P]UTP label incorporation in transcribed RNA, the shifted species in the pDR22A variant showed a strong radiolabeled band much greater than the radiolabel level in the shifted region for the C variant, which in turn was greater than that of the B variant, which was nearly at background levels (Fig. 3B). R-loop frequencies calculated by colony lift hybridization assay also showed that the pDR22A variant was the most efficient, with 5.1% of molecules in an R-loop conformation. In pDR22C, 1.2% of molecules had R-loops, which was an improvement over the B variant, in which only 0.6% of molecules were R-looped. The substantial difference (5.1% versus 0.6%) between pDR22A and pDR22B clearly indicates the G clusters in the RIZ are important for R-loop formation efficiency.

Even in the substrate with only 2G clusters in the REZ (pDR26A, pDR26C, and pDR26B), there was a small but discernible indication that G clustering in the RIZ improves R-loop formation. We could not detect any significant shifted species in the A, C, or B variants by ethidium staining or by radiolabel incorporation analysis, indicating that the REZ is not sufficiently capable of extending any R-loops initiated in the RIZ, presumably because of decreased G density of the transcript in the REZ (Fig. 4A and B). However, by colony lift hybridization, we detected one R-loop molecule out of 432 nontemplate strand derivative molecules (Table 1 and Fig. 4C). Also, the pDR54 set of substrates (Fig. 1) further support a role for G clusters in the RIZ in R-loop efficiency (Fig. 5). These substrates have a high G density in the REZ but no clustering. Compared to the substrate with motif B (0x4G) in the RIZ, motif A (2x4G), or motif C (1x4G) made R-loop formation more efficient (Fig. 5). Therefore, several sets of substrates (pDR22, pDR26, and pDR54 sets) suggest a role for G clustering in the RIZ.

FIG. 3. Effect of reducing the REZ G clusters from GGGG to GGG. Analysis and maps of R-loop molecules in pDR22A, pDR22C, and pDR22B with RIZ motif A (2×GGGG), C (1×GGGG), or B (no G clusters) and an identical REZ with maximum G-cluster size of GGG are shown. (A) Representation is similar to Fig. 2A. Linearized pDR22A, pDR22C, and pDR22B substrates were either mock transcribed (lanes 1, 6, and 11 for pDR22A, pDR22C, and pDR22B, respectively) or transcribed with T7 RNA polymerase in the presence of [α-32P]UTP and treated with RNase A afterward (lanes 2 to 4, lanes 6 to 8, and lanes 12 to 14 in triplicate for pDR22A, pDR22C, and pDR22B, respectively). The fifth lane in each set is transcribed sample treated with RNase A and RNase H1 (lanes 5, 10, and 15 for pDR22A, pDR22C, and pDR22B, respectively). The top panel is the ethidium-stained gel profile. The position of switch region containing linear fragment is designated “L.” The R-loop-induced shift is designated as “Shift.” (B) [α-32P]UTP radiolabel profile of the same gel shown in panel A after phosphorimagery exposure. The positions of shifted species and the linearized restriction fragment have been marked as “Shift” and “L,” respectively. (C) Representation of single-stranded regions in the DNA nontemplate strand detected by colony lift hybridization and sequencing after sodium bisulfite treatment. Similar to the description in Fig. 2C, the top line represents the linearized substrate, showing the T7 promoter, followed by the RIZ sequence A, C, or B upstream (shown as an inverted triangle) of the REZ that contains the modified switch repeats (GGG clusters) represented as thick arrows. The first line in each set shows all Cs on the nontemplate strand. Each of the following lines is an independent nontemplate strand derivative molecule, with vertical lines representing observed C-to-T conversions. The asterisks mark the position of the methylated C in bacterial dcm methylation sites (CCmecA/TGG) that remain unconverted.
Comparison of roles of G clusters versus G density in the R-loop initiation zone. In the studies described above, we observed that the addition of G clusters to the RIZ improves R-loop formation. We wondered whether initiation is strictly a function of the number of G clusters in the region immediately downstream of the promoter but upstream of the REZ and whether a sufficiently high G density in the RIZ can substitute for G clusters. The short motifs used in the previous section were too small to effectively disrupt the G clusters while maintaining the total G content. Therefore, we made a new substrate called pDR70 by inserting one repeat length of 50% methylated C in bacterial dcm methylation sites (CC\textsuperscript{met}A/TGG) that remain unconverted.

For this experiment, we call this the RIZ simply because it replaces the usual RIZ zone. The overall length of the switch substrate was maintained to be equivalent to four repeats of Sy3, but with the first repeat region containing no G clusters, and instead having a high G density due to alternating GNGNGN. Therefore, pDR70 was constructed to contain 24 Gs in the 48 nt representing the one repeat length of dispersed G region. For comparison, in pDR18 or pDR51, the nonrepeat strand G density in the first Sy3 repeat is 45.8% (22 Gs in the 48-nt repeat), most of which are in G clusters.

In the ethidium-stained gel, transcription induced R-loop formation was observed as a mobility-shifted species in substrates pDR51 and pDR18 (Fig. 6A and B, lanes 2, 3, and 4 and lanes 17, 18, and 19, respectively), which have G clusters in Sy3 repeats immediately downstream of the promoter but not in pDR70, which has the alternating G region followed by three Sy3 repeats (Fig. 6, lanes 13, 14, and 15). Comparison of the radiolabel densities at the shifted species shows that the pDR70 has ~17-fold less label intensity than that of the pDR18 (4 Sy3 repeats) and about 10-fold less compared to pDR51 (3 Sy3 repeats). This illustrates the effect of G clustering, even though there is a smaller effect of distance from the promoter. The values at the shifted position for pDR70 are comparable to pDR54, which has a four-repeat long G-rich region without G clusters (Fig. 6, lanes 7, 8, and 9). Both pDR70 and pDR18 have very similar and high G densities in the first repeat region and differ only in the distribution of the Gs within this region where R-loops would initiate.

These results show that G clustering is extremely important for efficient R-loop initiation regions. That is, replacing G clusters with unclustered, but equivalent G density sequences is inadequate and drastically reduces the efficiency of R-loop formation even in the presence of downstream G cluster-containing switch regions. Hence, we can conclude that high G density without clustering cannot replace the stronger effect of G clusters in R-loop initiation.

A high G density can compensate for an REZ with no G clustering. Although our studies above focused primarily on the RIZ, our substrates do have significant implication for clustering versus density in the REZ. Comparison of pDR54A with pDR26A is informative in this regard. Both substrates have the same RIZ. The REZ in pDR54A has only isolated Gs (no clusters of even 2 Gs), but it has a relatively high G density. The REZ in pDR26A has many GG clusters (Fig. 1).

We found efficient R-loop formation in pDR54A (4.2%) based on percentage shift values from the respective ethidium stained gels. In contrast, pDR26A has no detectable R-loop formation (compare pDR54A in Fig. 5A and B, lanes 2 to 4, with pDR26A in Fig. 4A and B, lanes 2 to 4). In fact, pDR54A is nearly as efficient in R-loop formation (4.2%) as pDR22A (5.3%), which has many GGG clusters throughout the REZ (Fig. 3A and B, lanes 2 to 4 for pDR22A and Fig. 5A and B, lanes 2 to 4 for pDR54A, respectively).

This clearly shows that a high G density in the REZ can support R-loop formation without depending upon G clusters. Therefore, a high G density in the REZ compensates for an REZ with no G clustering.
### Table 1. R-loop formation frequencies in various substrates calculated by colony lift hybridization assay

<table>
<thead>
<tr>
<th>Substrate</th>
<th>Initiation zone (RIZ)</th>
<th>Elongation zone (REZ)</th>
<th>No. of molecules with:</th>
<th>Frequency of R-loop formation (%)a</th>
</tr>
</thead>
<tbody>
<tr>
<td>pDR18A</td>
<td>2×4G clusters</td>
<td>Four Sy3 repeats (mostly GGGG clusters)</td>
<td>154</td>
<td>12.3</td>
</tr>
<tr>
<td>pDR18C</td>
<td>1×4G cluster</td>
<td>Four Sy3 repeats (mostly GGGG clusters)</td>
<td>220</td>
<td>5.9</td>
</tr>
<tr>
<td>pDR18B</td>
<td>No clusters</td>
<td>Four Sy3 repeats (mostly GGGG clusters)</td>
<td>371</td>
<td>0.8</td>
</tr>
<tr>
<td>pDR22A</td>
<td>2×4G clusters</td>
<td>Modified four Sy3 repeats (maximum cluster size is GGGG)</td>
<td>334</td>
<td>5.1</td>
</tr>
<tr>
<td>pDR22C</td>
<td>1×4G cluster</td>
<td>Modified four Sy3 repeats (maximum cluster size is GGGG)</td>
<td>411</td>
<td>1.2</td>
</tr>
<tr>
<td>pDR22B</td>
<td>No clusters</td>
<td>Modified four Sy3 repeats (maximum cluster size is GGGG)</td>
<td>1,051</td>
<td>0.6</td>
</tr>
<tr>
<td>pDR26A</td>
<td>2×4G clusters</td>
<td>Modified four Sy3 repeats (maximum cluster size is GGGG)</td>
<td>432</td>
<td>0.2</td>
</tr>
<tr>
<td>pDR26C</td>
<td>1×4G cluster</td>
<td>Modified four Sy3 repeats (maximum cluster size is GGGG)</td>
<td>754</td>
<td>&lt;0.1</td>
</tr>
<tr>
<td>pDR26B</td>
<td>No clusters</td>
<td>Modified four Sy3 repeats (maximum cluster size is GGGG)</td>
<td>1,066</td>
<td>ND</td>
</tr>
</tbody>
</table>

a That is, the percentage of molecules in the R-loop conformation.

b ND, none detected.

### DISCUSSION

R-loops have been studied in a wide range of in vitro and in vivo prokaryotic and eukaryotic (including mitochondrial) systems. However, mechanistic understanding of where and how they form has been lacking. In the previous in vitro or in vivo work on R-loops from our lab and others, no distinction was made between the zone where R-loops initiate (RIZ) and the remaining zone where the R-loop is maintained (REZ) (37). Although we previously learned a great deal about R-loops, including the threading back of the RNA for R-loop formation, we feel that the distinction between the RIZ and the REZ is critical for understanding where R-loops form and the length over which they extend. The sequence basis for the RIZ and REZ can be reduced to a thermodynamic level, which was not possible previously. In addition, definition of parameters for

![FIG. 5. Role of G clusters versus high G density in the RIZ in R-loop formation efficiency. Linearized pDR51 (three Sy3 G-clustered repeats), pDR54 (four-repeat long and dispersed high G density region without any G clustering), pDR70 (one repeat long dispersed and high G density region followed by three Sy3 repeats) and pDR18 (four Sy3 G-clustered repeats) were either mock transcribed (lanes 1, 6, 11, and 16 for pDR51, pDR54, pDR70, and pDR18, respectively) or transcribed with T7 RNA polymerase in the presence of [α-32P]UTP and treated with RNase A and RNase H1 (lanes 2 to 4, lanes 7 to 9, lanes 12 to 14, and lanes 17 to 19 in triplicates for pDR51, pDR54, pDR70, and pDR18, respectively). The fifth lane in each set is transcribed sample treated with RNase A and RNase H1 (lanes 5, 10, 15, and 20 for pDR51, pDR54, pDR70, and pDR18, respectively). The first repeat of these substrates (representing the RIZ) has approximately similar G density with dispersed (50% Gs in pDR54 and pDR70) or clustered (45.8% Gs in pDR51 and pDR18) distribution on the nontemplate strand. (A) The top panel is the ethidium-stained gel profile. The switch region containing linear fragment of pDR51 contains three repeats and therefore has a faster gel mobility than the switch region/modified switch region containing fragments of pDR54, pDR70, and pDR18, which have four repeat long regions. The positions of the linearized restriction fragments are marked as “Shift.” (B) [α-32P]UTP radiolabel profile of the same gel shown in panel A after phosphorimager exposure. The shifted position and the linearized restriction fragments have been marked as “Shift” and “L,” respectively. A concise description of the substrate switch regions is shown below panel B.
R-loop initiation and elongation now permits much more specific genome-wide searches for potential R-loop forming regions. Assaying for R-loop forming regions requires estimates not only of where they might initiate but also of how far downstream they might extend.

**Mechanism of R-loop initiation.** The studies here markedly improve our understanding of the mechanism of R-loop formation (Fig. 7). During transcription of random sequence by all RNA polymerases, the nontemplate DNA strand is separated from the template DNA strand. The nontemplate DNA strand appears to track along the outside of the RNA polymerase for a length because it is solvent exposed, is susceptible to nucleases, and can be recognized by single-stranded binding proteins (2, 3, 43). The two DNA strands appear to reanneal outside of the upstream side of the RNA polymerase or on the surface of the polymerase in a region where both DNA strands are solvent exposed (23, 44). It is clear that the nascent RNA that exits the RNA polymerase is single stranded, based on its susceptibility to RNase A and T1 (37). R-loop formation upon transcription by phage and mammalian polymerases indicates that the nascent RNA strand can compete with the nontemplate DNA strand for annealing to the template DNA strand as it emerges from the RNA polymerase.

The competition between the RNA and the nontemplate DNA strand for annealing to the template DNA does not have to occur at the instant that all three strands exit the RNA polymerase. Rather, the two DNA strands might anneal to one another at or close to the surface of the RNA polymerase and then breathe open, thereby providing an opportunity for a G cluster of the RNA strand to invade the DNA. Such breathing of the newly reformed DNA duplex is particularly likely because as the two DNA strands anneal to one another on the upstream side of the RNA polymerase, the newly reannealed DNA duplex at this position is effectively a DNA end, and DNA ends are known to undergo substantial breathing (41). Interestingly, breathing of G-rich sequences is maximal when one strand is composed of consecutive Gs and the other is composed of consecutive Cs, just as is the case for the G clusters in the RIZ (41). This may be because such GGG/CCC (or longer) regions adopt a DNA conformation that is intermediate between the B form and the A form, called B/A-intermediate DNA, and the B/A-intermediate conformation is favorable for DNA breathing (41).

All RNA-DNA is stronger than DNA-DNA of the corresponding sequence. However, the resumption of DNA-DNA in random DNA sequences is likely because the nontemplate strand is closer to the template strand than the RNA strand. Even for linear Ig switch region sequences with 50% G density in clusters, only a small minority of substrates assume an R-loop conformation. Therefore, despite the thermodynamic advantage of RNA-DNA over DNA-DNA, the proximity advantage of the nontemplate DNA strand is a dominant factor in favoring DNA-DNA over RNA-DNA. Therefore, the initial nucleation site of the thread-back RNA must have maximum stability via as many Gs as possible in a short length (because there is not sufficient length of DNA strand separation to permit a long segment of RNA to bind). It is interesting in this regard that most of the R-loop substrates with two G clusters initiate within that two G cluster RIZ, whereas almost all but one R-loop molecule in substrates with zero RIZ G clusters started downstream of the RIZ, mostly at or inside the switch regions (REZ). Although the influence of the one G cluster motif in the RIZ results in increased R-loop formation compared to the zero G-cluster substrates, the R-loop start sites on these substrates are more varied, pointing to the intermediate stability of R-loops initiated at a one G-cluster RIZ.

When we compare the sequence between pDR18A and
pDR18B, the two substrates differ by two extra G clusters in RIZ of pDR18A not present in pDR18B. This is a comparatively small change in the overall sequence. However, the R-loop formation efficiency of pDR18A is ~15-fold greater than pDR18B. This disproportionate increase in pDR18A is observed because the 2x4G clusters in pDR18A are closer to the 5’ end of the transcript, whereas the position of the first G cluster in pDR18B is internal in the transcript. The transcript length is the same in the two substrates. Due to the higher mobility of the nucleotide positions near the 5’ terminus compared to the internal positions, the 5’ ends of the transcripts have a higher collision frequency with the template DNA strand. Thus, G clusters present toward the free 5’ end of the RNA transcript have a higher probability of nucleating an R-loop initiation event. Once initiated, the overall stability of the R-loop is also higher in the A variant because of a higher local G-content around this region (more G clusters in the pDR18A compared to pDR18B; see Fig. 1), thereby further reducing the RNA dissociation propensity (Fig. 7). Thus, whereas the stability factors (total G-content) are additive in nature, the R-loop initiation factors (increased molecular mobility driving higher collision frequency and RNA-DNA nucleation upstream of the polymerase) are much more additive.

**Thermodynamic considerations.** It is useful to consider the thermodynamic aspects to fully appreciate the mechanism of RNA-DNA hybrid formation in R-loops. We observe a many fold improvement in R-looping when the RIZ contains one or two clusters of GGGG. This improved clustering correlates much better with the stability of annealing in the RIZ than in the entire R-loop region (RIZ + REZ). The efficiency of R-looping increases as the length and number of G clusters increase. As mentioned earlier, all RNA-DNA duplexes are stronger than DNA-DNA duplexes (38; http://ozone3.chem.wayne.edu). The addition of a motif containing two GGGG clusters or one GGGG cluster can be calculated to improve the thermodynamic stability of DNA-DNA that has clusters of Gs is weaker than DNA-DNA and DNA-DNA gets smaller. At this point, the proximity advantage for the annealing of the nontemplate DNA to the template DNA eventually prevails, and the R-loop elongation ends.

The thermodynamic stability of clustered G duplexes in a DNA duplex is weaker than that of dispersed sequences such as GNGNGN. The clustered G duplexes may have better intrastand base stacking at the expense of interstrand interaction. A greater tendency of these G-clustered sequences to breathe (as mentioned above) may also make the DNA-DNA component weaker. The majority of the difference between the RNA-DNA and DNA-DNA components is contributed by this DNA-DNA interaction effect rather than the RNA-DNA interaction within zones that contain G clusters.

**Relevance of R-loop initiation and elongation zones in vivo.** In light of our inferences here, one might wonder why mammalian switch regions evolved to have G clustering throughout, rather than only at the beginning. R-loop initiation is a stochastic process and is not 100% efficient at the first G cluster. In fact, on linear substrates with four 5’ G4G repeats, less than 10% of the molecules are in an R-loop conformation at any one time. Hence, additional G clusters further downstream improve the R-loop formation efficiency overall. Therefore, it is not surprising that the Ig switch regions contain G-clustered repeats throughout their repetitive zone. Mapping of R-loop positions in vivo shows that the initiation point varies considerably (47), a finding consistent with what we see in vitro for partial switch regions (37).

The findings here provide a basis for understanding the R-loop initiation and extension seen upstream of the Igµ switch region in vivo (17). A strong RIZ followed downstream by a very weak REZ may be insufficient to remain stable as an R-loop. An in vivo example of this may be the 50-bp G-clustered (50% G-dense on the nontemplate strand) peak upstream of the Sµ repetitive region. This might be insufficient to initiate stable R-looping in the wild-type allele or an allele that deletes the core Sµ repeats (called ΔSµ/TR in reference 17) if not for the region downstream of it (REZ), which is G-dense but relatively unclustered. Therefore, the observations here for R-loop initiation and elongation are likely to have predictive value in assessing transcription units for their propensity for R-loop initiation and elongation.

**ACKNOWLEDGMENTS**

We thank Chih-Lin Hsieh and members of the Lieber laboratory for discussion. We thank Albert Tsai of our lab for the use of the bisulfite plottor for the R-loop maps.

This study was supported by the National Institutes of Health (M.R.L.). D.R. is a recipient of a USC Hiedelberger Award and a CBM Training Grant Award.

**REFERENCES**


