DNA Repair xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

DNA Repair

journal homepage: www.elsevier.com/locate/dnarepair

The contribution of co-transcriptional RNA:DNA hybrid structures to DNA damage and genome instability

Stephan Hamperl, Karlene A. Cimprich*

Department of Chemical, Systems Biology, Stanford University School of Medicine, 318 Campus Drive, Stanford, CA 94305-5441, USA

ARTICLE INFO

Article history:
Available online xxx

Keywords:
R-loops
Genome instability
Double-strand breaks
G-quadruplex
Topoisomerase
THO/TREX complex
THSC/TREX-2
ASF splicing factor
mRNP biogenesis
RNA processing factors
RNase H
RNA:DNA helicases
Senataxin
Activation-induced deaminase (AID)
APOBEC family
Transcription–replication conflicts

ABSTRACT

Accurate DNA replication and DNA repair are crucial for the maintenance of genome stability, and it is generally accepted that failure of these processes is a major source of DNA damage in cells. Intriguingly, recent evidence suggests that DNA damage is more likely to occur at genomic loci with high transcriptional activity. Furthermore, loss of certain RNA processing factors in eukaryotic cells is associated with increased formation of co-transcriptional RNA:DNA hybrid structures known as R-loops, resulting in double-strand breaks (DSBs) and DNA damage. However, the molecular mechanisms by which R-loop structures ultimately lead to DNA breaks and genome instability is not well understood. In this review, we summarize the current knowledge about the formation, recognition and processing of RNA:DNA hybrids, and discuss possible mechanisms by which these structures contribute to DNA damage and genome instability in the cell.

1. Introduction

RNA synthesis is one of the central processes by which other cellular machineries access the genetic information encoded by genomic DNA. Therefore, transcription represents an essential process of DNA metabolism. However, actively transcribing RNA polymerases (RNAPs) inevitably induce fundamental alterations in the underlying chromatin template in eukaryotic cells. The tight association of nucleosomes with the template DNA must be disrupted during transcription elongation, leading to partial or complete loss or exchange of histone molecules [1–6]. Moreover, when unwinding the DNA double helix at the site of active transcription, negative and positive supercoils arise behind and in front of the advancing RNAP, respectively [7,8]. These rearrangements have potentially destabilizing consequences for the underlying DNA molecule. Not surprisingly, highly transcribed genes exhibit increased mutation and recombination rates compared to genomic regions with low transcriptional activity [9–12]. These transcription-associated mutagenesis (TAM) and transcription-associated recombination (TAR) events are conserved from bacteria to mammalian cells, and extensive research in the past few years has led to great progress in our understanding of the molecular mechanisms leading to TAR and TAM. A large body of evidence suggests that conflicts between the transcription and replication machineries are a major source of the observed genomic instabilities, and we refer the reader to several excellent reviews covering this topic [13–19]. However, work in the last few years suggests that formation of co-transcriptional RNA:DNA hybrid structures, known as R-loops, may significantly contribute to the above phenomena. In an R-loop structure, the nascent RNA strand invades the DNA duplex to hybridize with the complementary template strand after it exits RNAP. This results in a nucleic acid structure containing

Abbreviations: DSB(s), double-strand break(s); RNAP(s), RNA polymerase(s); TAM, transcription-associated mutagenesis; TAR, transcription-associated recombination; ssDNA, single-stranded DNA; CSR, class switch recombination; AID, activation induced deaminase; SSB(s), single-strand break(s); rDNA, ribosomal DNA; ChIP, chromatin immuno-precipitation; RNP, ribonucleoprotein; CIN, chromosome instability; CTD, carboxy-terminal domain; CFS, common fragile site; APOBEC, apolipoprotein B mRNA-editing catalytic poly-peptide.

* Corresponding author. Tel.: +1 650 498 4720; fax: +1 650 725 4665.
E-mail address: cimprich@stanford.edu (K.A. Cimprich).

http://dx.doi.org/10.1016/j.dnarep.2014.03.023
1568-7864/

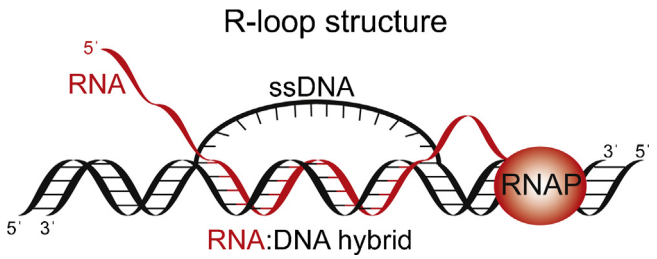

Fig. 1. Schematic representation of an R-loop structure.
The cartoon depicts the general structure of an R-loop. The nascent RNA strand (red) is synthesized by RNA polymerase (RNAP, red oval) and hybridizes with the complementary DNA template strand. The non-template strand is exposed as single-stranded DNA (ssDNA).

an RNA:DNA hybrid and a displaced tract of single-stranded DNA (ssDNA) [13,20,21] (Fig. 1).

RNA:DNA hybrids are implicated in a multitude of biological processes. Besides the short RNA primers of ~7–12 bp generated by DNA primase during replication of the lagging strand [22] and transient formation at the center of the transcription bubble (~8 bp) during transcription by RNAPII [23], longer and more stable stretches of hybrids are key intermediates in replication and recombination. For example in *Escherichia coli*, replication initiation of *ColE1*-type plasmids requires transcription-dependent formation of a stable RNA:DNA hybrid that extends past the origin of replication. Cleavage of the hybrid by RNase H, which specifically degrades the RNA in an RNA:DNA-hybrid, leaves a 3'—OH end that is extended by the replication machinery [24]. A similar mechanism has been proposed for replication of yeast and mammalian mitochondrial DNA, with DNA synthesis being primed by a mitochondrial RNAP transcript also processed by RNase H [25–27]. Finally, the formation of RNA:DNA hybrid structures plays a role in the generation of antibody diversity during class switch recombination (CSR) in activated B cells (reviewed in [28,29]). In this process, transcripts derived from the repetitive switch (S) sequences of IgH genes form an R-loop with the template strand *in vitro* and *in vivo* [30–35]. The ssDNA of these R-loops can then be targeted by activation-induced deaminase (AID) [36], and the resulting deoxyuridine residues are processed by components of the base-excision or mismatch repair machineries to single-strand breaks (SSBs) [37–40]. These DNA lesions are finally converted to a DSB, a necessary intermediate for recombination at the S sequences, in a process that involves non-homologous end joining factors [41,42].

The events occurring during CSR clearly highlight the potential for co-transcriptional formation of RNA:DNA hybrid structures to induce DNA breaks and recombination. They also raise the possibility that CSR-related mechanisms could contribute to R-loop mediated strand-break formation and chromosomal instability at other genomic regions and in other cell types. In this regard, numerous recent studies suggest that R-loops may form with higher frequency in eukaryotic genomes than previously anticipated. Immunofluorescence experiments performed using an antibody which detects RNA:DNA hybrids in a sequence-independent manner [43] showed abundant signals distributed throughout the nucleoplasm in human H1 ESC and mouse NPC cells [44,45]. Moreover, DNA:RNA immunoprecipitation (DRIP) combined with high-throughput sequencing (DRIP-seq) detected putative RNA:DNA hybrids at more than 20,000 peak regions in human Ntera2 cells [44]. A recent bioinformatic study corroborated these results by creating a computational algorithm to identify potential R-loop forming sequences (RLFS) in the human genome. Strikingly, almost 60% of transcribed sequences contained at least 1 RLFS [46]. Thus, R-loops may be abundant cellular intermediates. This finding suggests that cells may also have systems that prevent the processing of these R-loops into DNA breaks. Here, we summarize the current knowledge of the factors and cellular pathways implicated in the formation, recognition and processing of R-loop structures *in vivo*. Finally, we discuss possible mechanisms for how these aberrant nucleic acid structures may lead to DNA breaks and compromise genome integrity.

2. Formation of R-loops is dependent on the sequence and transcription-dependent topological state of the DNA molecule

2.1. Clusters of G-rich sequences and G-quadruplex structures

The molecular mechanism of R-loop formation has primarily been elucidated from *in vitro* transcription experiments which utilized prokaryotic or phage RNA polymerases and purified plasmid DNA coding for mammalian class-switch regions (see Section 1) [30,33,34,47–49]. In an elegant set of experiments, Roy and colleagues showed that the nascent RNA strand must pass through the exit pore of RNA polymerase before threading back to anneal with the template DNA (thread-back model), and that therefore the R-loop is not just an extension of the ~8 bp hybrid formed in the transcription bubble (extended hybrid model) [49]. This is consistent with the conserved architecture of all cellular RNA polymerases, which requires that RNA and DNA strands exit at different sites from the enzyme [50–55].

The process of R-loop formation necessitates a competition between the nascent RNA and the non-template DNA strand to hybridize with the template strand. Therefore, hybrid formation should be thermodynamically favorable when compared to reannealing of the DNA duplex in the R-loop forming region. Indeed, synthetic RNA:DNA hybrid structures with a high RNA-purine/DNA-pyrimidine ratio were shown to be more stable than a DNA:DNA duplex of the same sequence composition [56–58]. Thus, a high guanine (G) density in the non-template DNA strand promotes R-loop formation *in vitro* and *in vivo* [35,49]. More precisely, one or two clusters of consecutive (3 or more) G residues in the R-loop initiating zone efficiently nucleates hybrid formation, whereas a high G density (but not G clustering) is sufficient for elongation of the R-loop (Fig. 2A, [59]).

Other factors on the non-template strand may also drive R-loop formation. A recent study showed that nicks in the non-template DNA strand reduce its ability to reanneal to the template strand, thereby favoring hybridization of the RNA and R-loop formation, even in the absence of G clusters [60]. As nicks are created frequently in the genome by exogenous or other endogenous sources of DNA damage, this finding may further expand the possible R-loop forming regions *in vivo*. Also on the non-template strand, clusters of G-rich sequences have the potential to fold into a secondary, non-B-form DNA structure referred to as G-quadruplex or G4 DNA (reviewed in [61]). G4 DNA is characterized by the association of four guanines bound through Hoogsteen base pairing and variable stacks of guanine quartet planes (Fig. 2B, [62]). Intriguingly, G-quadruplex structures have been directly observed during *in vitro* transcription of S regions by electron microscopy [47], and were recently shown to be stable structures, detectable by immunostaining and sequencing analysis, in human genomic DNA [63,64]. The high stability of G4 DNA may help stabilize the single-stranded tract of DNA in an R-loop structure, making it tempting to speculate that formation of G4 DNA on the non-template strand may facilitate or contribute to RNA:DNA hybrid formation during transcription *in vivo* (Fig. 2B).

Please cite this article in press as: S. Hamperl, K.A. Cimprich, The contribution of co-transcriptional RNA:DNA hybrid structures to DNA damage and genome instability, DNA Repair (2014), http://dx.doi.org/10.1016/j.dnarep.2014.03.023