1. **RNA:DNA Hybrids**

1.1 RNA:DNA hybrids as beneficial structures. What are Hybrids?

RNA:DNA hybrids form upon RNA annealing to the DNA template strand or any homologous DNA region (Westover 2004, Wahba 2013, Cloutier 2016, Ariel 2020). Transient RNA:DNA pairs are typically required to prime DNA replication and regulate gene expression (Nadel 2015). More long and stable RNA:DNA heteroduplex can force the non-template DNA strand into a looped-out state, resulting in an R-loop structure (Malig 2020). The helical conformation of the R-loops differs from the conventional right-handed DNA double-helical structure (B-DNA), and it is referred to as non-B DNA. Recent works (Malig 2020, Chedin & Benham 2020) unveiled that R-loops are the largest of non-B DNA structures, and are specifically fit to absorb large amounts of negative DNA supercoiling (see below).

R-loops primarily occur co-trascriptionally (*cis*), upon the hybridization of the nascent RNA with the template DNA strand from which it was transcribed, behind the advancing RNA polymerase (RNAP) (Westover *et al.*, 2004). Some findings suggested that post-transcriptional R-loops might arise in *trans,* upon the invasion of a non-coding RNA (ncRNA) into complementary double-stranded DNA (dsDNA), other than that from which it was synthesized (Wahba et al, 2013; Cloutier et al, 2016; Ariel et al, 2020). In this context, a single RNA species may affect numerous genome loci with similar sequence motifs, such as repeated and satellite elements (Wahba et al. 2013). This is particularly the case for the centromeric R-loops, favoured by a trans-activating ncRNA, annealing the CENP-B DNA binding sites, in centromeric repeat-dense satellites across chromosomes (Masumoto et al. 1989). Additional protein-mediated catalysis, such as homology-directed DNA recombination machinery (Zaitsev 2000) or CRISPR-Cas systems (Xiao 2017, Jiang 2016), is required to promote the R-loop initiation in *trans*.

Many factors contribute to R-loop formation and stability. R-loop initiation may be affected by DNA base-pairing energetics and topology. Generally, RNA:DNA base pairs feature lower energy compared to DNA:DNA base pairs, typically in G-rich and GC-skewed regions. Indeed, evidence from experimental and mapping studies showed that R-loops hotspots are generally enriched in G-rich and G/A-rich regions of transcripts, due to the thermodynamic stability of riboG:deoxyC in the RNA:DNA hybrids (Huppert 2008, Ratmeyer 1994). Repetitive regions featuring GC- (Ginno 2012, Ginno 2013) and AT-skew (Wahba 2016), G-rich (De Magis 2019) and tandem repeats (Su 2017, Groh 2014) are more inclined to R-loops formation. In this respect, R-loops support genome stability by promoting chromatin condensation or chromosome segregation in repeat-dense centromeric regions (Castellano-Pozo Mol. cell 2013, Kabeche 2018), and forming telomere-repairing R-loops in telomere repeat-containing TERRA (Graf 2017).

On the other hand, an increased negative superhelicity may model the frequency and the distribution of R-loops, even over unfavourable DNA regions. Negative torsional tension constitutes a high-energy and stressed state of the DNA. Along this line, R-loop initiation may efficiently mitigate such stressed state, by absorbing the local under twist and relaxing nearby domains (Stolz 2019). Good evidence of this was found in human cells, where depletion of DNA topoisomerase I led to the R-loop accumulation in long, highly transcribed, and physically constrained genes (Manzo 2018). Analogous results were obtained from SMRF-Seq experiments, profiling R-loops in unfavourable sequences at gene bodies or terminal genic regions, probably affected by local superhelical stress (Malig 2020, Chedin & Benham 2020). By contrast, regions with the most favourable RNA:DNA energetics, including highly GC-skewed CpG island promoters, enable R-loops at low levels of superhelical stress (Stolz 2019).

Under most conditions, R-loop formation is the product of a balance between DNA sequence and topology.

The R-loop forms may be further stabilized when the displaced strand partially folds into a G-quadruplex (G4) structure (Manzo 2018, Chedin 2016). In eukaryotes, G4s are four-stranded non-B DNA structures usually enriched at replication origins, gene promoters, untranslated exon regions, microsatellite and telomeric repeats (Yang 2019). Notably, Zhang et al. (Zhang 2019) demonstrated that R-loops might affect G4 accumulation at telomeres, showing a decrease in telomere G4 in response to TERRA depletion.

G4s and R-loops can assemble at the same time at highly active genes in cells (Brambati 2020, Belotserkovskii 2018), generating structures known as G-loops (Duquette 2004). Similar to R-loops, G-loop initiation relies on transcription rate, non-template strand G-richness, RNA:DNA stability, and negative supercoiling of the template DNA.

2. R-loop detection on repeats in cancer cell lines

2.1 The methodology for the detection of RNA:DNA hybrids, advantages and disadvantages, discrepancies and result interpretation

Commonly used technologies for genome-wide R-Loop mapping have been summarised in recent reviews, along with a discussion of their highlights and challenges (García-Muse and Aguilera 2019, Chedin 2021, Crossley 2019, Vanoosthuyse 2018). Up to date, explorative R-loops maps have been obtained employing DNA:RNA immunoprecipitation combined with sequencing (DRIP-seq) (Ginno 2012), R-loop chromatin enrichment (R-ChIP) (Chen 2019, Yan 2019), and more recently, single-molecule R-loop footprinting (SMRF-seq) methods (Malig 2020).

DRIP-seq strategies locate R-loops by sequencing DNA or RNA signals of S9.6 immunoprecipitated RNA:DNA hybrids. Collectively these DRIP-based studies allowed the capture of R-loop signals on transcribed gene bodies, CpG island promoters and terminal genic regions (Sanz 2016). Although many modifications of these protocols have been implemented (for example, the strand-specific DRIPc- (Sanz 2016-2019, Hartono 2018), bisDRIP- (Dumelie 2017), ssDRIP-seq (Xu 2017) methods), they suffer from inconsistency in R-loop locations and numbers. This variance in results can be explained by differences in DNA extraction methods, cell cycle profiles of starting material, and antibody specificity (Pan 2020, Vanoosthuyse 2018, Hartono 2018).

Alternative approaches employ different catalytically inactive RNase H enzymes to recognise RNA:DNA heteroduplexes, and include among others, DRIVE (DNA:RNA in vitro enrichment) (Ginno 2012), R-ChIP (R-loop chromatin enrichment) (Chen 2019) and MapR strategy (Yan 2020, Yan 2019), founded on the principles of CUT&RUN. R-ChIP-based studies globally detected a smaller signal number, but were more sensitive to transient and also smaller R-loops located at i) G-rich loci associated with promoter-proximal pausing of RNA polymerase II, ii) active enhancers (Wulfridge 2021), and iii) tRNAs, compared to DRIP-seq signals, also in human cells (Chen 2017, Legros 2014, Roy 2009). Anyway, the generation of stable cell lines expressing a catalytic mutant RNase H1 hampers the adaptability of these methods, mostly in mammalian cells (Sanz 2019, Chen 2017, Sanz 2016). Furthermore, over-expression of RNase H1 in human cells has been shown to impact R-loop homeostasis by affecting cell transcriptome and proteome (Kabeche 2018).

A long-read and single-molecule sequencing technology, the single-molecule R-loop footprinting (SMRF-seq), has been recently fulfilled but it is not broadly used yet. It is based on the reactivity of ssDNA to bisulfite-mediated cytosine deamination, to characterize strand-specific R-loops on single-molecule amplicon, at ultra-deep coverage (Malig 2020, Chedin 2021). This strategy has shown R-loop formation through gene bodies and over terminal regions, in agreement with bulk DRIP-based studies (Malig 2020).

Pearson's correlation analysis performed by the Chedin group, on signals from a large collection of existing human cell datasets, demonstrated more consistent endings from classical DRIP-based experiments with respect to others (Chedin 2021).

Besides the identification of such critical issues, attempts to develop standardized experimental and computational procedures have been made, to make the produced data comparable. Thus, guidelines for best practices when working with R-loops have been postulated in (Vanoosthuyse 2018, Sanz & Chedin, 2019) and (Chedin 2021) references. Best practices for high-quality results, when performing DRIP-seq experiments and data interpretation, include: i) the use of at least two independent biological replicates; ii) the use of spike-in control to normalize sequencing data (Crossley 2020); iii) the validation of a peak subset with an alternative method, for example, DRIP-qPCR; iv) the implementation of the RNAse H-treated controls; v) the comparison of obtained data with high-quality, already published datasets, as done in Chedin 2021, iv) also by visually inspection in track-based views aligned to reference genome, i.e. UCSC Genome Browser (Kent 2020) or IGV (Thorvaldsdóttir 2013). Moreover, when mapping R-loops via RNA, signals abnormally enriched over exons and short interspersed elements (SINEs) such as Alu elements, should be indicative of RNA contaminations (Chedin 2021). Lastly, much emphasis has been placed on the use of long-read sequencing technologies, potentially more effective for investigations of R-loop formation in repetitive genome regions (Chedin 2021, Vanoosthuyse 2018).

In this regard, previous studies outlined several examples in which R-loop formation was associated with low-complexity and simple repeats among species (Zeng 2021, Yan 2019, Velazquez Camacho 2019, Johnson 2017). Based on these experimental findings, Wongsurawat and colleagues designed an in silico R-loop predictive model, considering G-cluster regions important for an efficient R-loop initiation (Wongsurawat 2011). However, the comprehensive understanding of the R-loop-repeats association is currently undermined by the incomplete annotation of highly variable repeat-dense regions, such as rDNA loci, TEs, satellites centromeres, pericentric regions and telomeres in genome reference sequences (Altemose 2014; Rosenbloom 2015). Accordingly, while standard reference-based mapping tools (Lee 2018) are not useful for these regions, reference independent methods have been implemented to explore the genome for repeat content (reviewed in O'Neill 2020). For example, repeat masking pipelines can be used to annotate types of repeats and their frequency within a given high-throughput sequencing dataset (Bailly-Bechet 2014). If a particular repeat class is known, K-mer based approaches can also be used to classify reads into specific repeat groups (see, for example, Cong Feng 2020). On these grounds, the support of long-read sequencing technologies to the complete telomere-to-telomere genome assembly (Miga 2020), promises new insights also into the characterization of repeat-dense R-loop-prone regions as well (Vanoosthuyse 2018, Chedin 2021).

To sum up, knowledge regarding R-loop physiological meaning and distribution in cells is a function of the cell cycle stage, growth conditions, transcription and epigenetic profiles, and DNA sequence features, that can affect the detection methods at the molecular level, especially in complex high-repetitive sequences (Garcia and Muse 2019).

2.2 R- loops mapping across the genome: within promoter regions and repetitive sequences short part on "normal" regions (promoters, classic transcription units)

Although globally descriptive research has been carried out on R-loops, few studies exist which adequately covers R-loop association with repeat sequence. A serious weakness with this argument is that such repetitive regions, including telomeres and centromeres, are poorly accessible by short-read sequencing technologies, as described above.

Studies from many laboratories unveiled predominant R-loops at gene promoter regions harbouring CpG islands, and transcription termination sites enriched with GC-skews. Strong evidence of R-loop sequence preference for G/C content (Chambers et al., 2015), and G/C skew within gene promoters was found by Chen et al. (Chen 2017), in embryonic kidney HEK293T and erythroleukemia K562 cells, after they developed the in vivo RNase H-based strategy for R-loop profiling. These findings suggest that R-loop levels are generally dependent on G/C content.

This R-loop feature seems to be a commonality in several human cancer cells, including cervical cancer HeLa (Tan-Wong 2019, Hamperl 2017, Promonet 2020), embryonal carcinoma NTera2 (Sanz 2016), osteosarcoma U2 OS (Cohen 2018, Gorthi 2018), Ewing sarcoma CHLA10, EWS502, TC32 (Gorthi 2018), and erythroleukemia K562 cells (Sanz 2016, Chen 2017).

Up to date, several studies have reported the relevance of the R-loops as common vulnerabilities in cancers (Nguyen 2019, Boros-Oláh 2019, Patel 2021), also in relation to their sequence features. A study from De Magis et al. (De Magis 2019) reports the G-quadruplexes contributing to the R-loop stability at G-rich loci in promoters or terminators. These structures are involved in the regulation of different biological pathways, as well as genome instability (Maffia 2020). On these bases, G-quadruplex-targeting drugs revealed a mechanism of cell killing by G4 ligand-induced DNA damage through unscheduled G4/R-loop structures, in BRCA2 defective cancer cells (De Magis 2019). Consistently with these findings, further data collection and research are required to determine exactly how R-loops are linked to repetitive loci, also in response to pharmacological screens in the cancer field.

Wu et al (Wu et al. 2020) studied the role of RTEL1 (regulator of telomere elongation helicase 1) in promoting the dissolution of R-loops, in repeats including common fragile sites (CFS), rDNA and telomeres regions, more prone to accumulate G-quadruplex-associated R-loops during S phase (Reddy 2014). After RTEL1 depletion and aphidicolin (inhibitor of eukaryotic DNA replication) treatment, they found an increase of R-loops accumulation in U2OS, HeLa and colon cancer HCT116 cells. The most relevant finding to emerge from this study was that R-loop were particularly enriched within CFS regions, and generally associated with transcribed genes and multiple G4-rich sequences. Importantly, the authors observed a large proportion of the peaks that overlapped with cancer-associated mutations (Tate 2019), in line with other studies reporting the CFS instability at the early stage of cancer development, as hotspots for chromosomal rearrangements (Glover 2017).

A more recent study by Zeng et al. (Zeng 2021) focuses on the R-loop-repeats association across diverse species. The authors disclosed that about 20% of the DRIP-seq signals contained repetitive elements. Among these, transposable elements, such as LINEs, LTRs, short interspersed nuclear elements (SINEs) and DNA families have poorly represented in osteosarcoma U2OS cells, while rRNAs were mildly enriched. A strong positive correlation between cis R-loop formation and retroposons and satellites were instead observed in human U2 OS. These findings confirmed the results previously obtained in (Yan 2019, Johnson 2017, Sanz 2016, Velaskez-Comacho 2020), proposing repetitive elements as key features of R-loop formation in promoter regions.

Moran et al. (Moran 2021) reported the characterization of R-loop formation during mitosis. In this study, mitotic DRIP-seq peaks were compared to those from asynchronous HEK293 cells, already published by Nadel et al. (Nadel et al., 2015). The authors showed a decrease of R-loop at promoters and genes, and an R-loop accumulation in repeat elements in mitosis, compared to interphase cells. Another major finding was that mitotic R-loops were distinct from interphase ones, and were enriched at alpha-satellite repeats (ALR), scaffold attachment repeats (SAR), and Human satellite II repeats (HSATII). As mitotic R-loop content was further increased after Aurora B inhibition, the authors concluded that Aurora B activity was required to deplete these R-loops, consistent with the roles of Aurora B in centromeric regulation and cohesion (Kabeche et 2018). The authors conclusively explained that R-loops might form during chromosome condensation across chromosomes but are restricted to centromeres after prophase and are preferentially enriched within repetitive sequences, including centric and pericentric repeats.

Finally, Sagie et al. (Sagie 2017) carried out a DRIP-seq analysis on human primary fibroblasts and an embryonal carcinoma cell line, unveiling many subtelomeres, and in particular those that display a strong positive GC skew, enriched for R-loops.

Considerably, much more systematic work will need to be done to determine an unambiguous and functional relationship between R-loops and repetitive DNA elements in cancer.

References

Aguilera, A. & García-Muse, T. R loops: from transcription byproducts to threats to genome stability. Mol. cell 46, 115–124 (2012).

Altemose N, Miga KH, Maggioni M, Willard HF (2014) Genomic characterization of large heterochromatic gaps in the human genome assembly PLoS Comput Biol

Bailly-Bechet M. et al. (2014) “one code to find them all”: a perl tool to conveniently parse repeatmasker output files. Mobile DNA, 5, 13.

Brambati A, Zardoni L, Nardini E, Pellicioli A, Liberi G. The dark side of RNA:DNA hybrids. Mutat Res. 2020 Apr-Jun;784:108300. doi: 10.1016/j.mrrev.2020.108300. Epub 2020 Feb 29. PMID: 32430097.

Castellano-Pozo, M. et al. R loops are linked to histone h3 s10 phosphorylation and chromatin condensation. Mol. cell 52, 583–590 (2013).

Chédin F, Hartono SR, Sanz LA, Vanoosthuyse V. Best practices for the visualization, mapping, and manipulation of R-loops. EMBO J. 2021 Feb 15;40(4):e106394.

Chédin, F. Nascent connections: R-loops and chromatin patterning. Trends Genet. 32, 828–838 (2016).

Chen, L.; Chen, J.Y.; Zhang, X.; Gu, Y.; Xiao, R.; Shao, C.; Tang, P.; Qian, H.; Luo, D.; Li, H.; et al. R-ChIPusing inactive RNase H reveals dynamic coupling of R-loops with transcriptional pausing at gene promoters.Mol. Cell2017,68, 745–757.e5.

Cong Feng, Min Dai, Yongjing Liu, Ming Chen, Sequence repetitiveness quantification and de novo repeat detection by weighted k-mer coverage, Briefings in Bioinformatics, 2020

Crossley, M. P., Bocek, M. & Cimprich, K. A. R-loops as cellular regulators and genomic threats. Mol. cell 73, 398–411 (2019).

Crossley, M. P., Bocek, M. J., Hamperl, S., Swigut, T. & Cimprich, K. A. qdrip: a method to quantitatively assess rna–dna hybrid formation genome-wide. Nucleic Acids Res. (2020).

De Magis, A. et al. Dna damage and genome instability by g-quadruplex ligands are mediated by r loops in human cancer cells. Proc. Natl. Acad. Sci. 116, 816–825 (2019).

Dumelie, J.G.; Jaffrey, S.R. Defining the location of promoter-associated R-loops at near-nucleotide resolutionusing bisDRIP-seq.eLife2017,6, e28306. [CrossRef] [PubMed]

García-Muse, T. & Aguilera, A. R loops: from physiological to pathological roles. Cell 179, 604–618 (2019).

Ginno, P. A., Lim, Y. W., Lott, P. L., Korf, I. & Chédin, F. Gc skew at the 50 and 30 ends of human genes links r-loop formation to epigenetic regulation and transcription termination. Genome research 23, 1590–1600 (2013).

Ginno, P. A., Lott, P. L., Christensen, H. C., Korf, I. & Chédin, F. R-loop formation is a distinctive characteristic of unmethylated human cpg island promoters. Mol. cell 45, 814–825 (2012).

Graf, M. et al. Telomere length determines terra and r-loop regulation through the cell cycle. Cell 170, 72–85 (2017).

Grunseich, C. et al. Senataxin mutation reveals how r-loops promote transcription by blocking dna methylation at gene promoters. Mol. cell 69, 426–437 (2018).

Hartono, S.R.; Malapert, A.; Legros, P.; Bernard, P.; Chedin, F.; Vanoosthuyse, V. The affinity of the S9.6antibody for double-stranded RNAs impacts the accurate mapping of R-loops in fission yeast.J. Mol. Biol.2018,430, 272–284

Johnson WL, Yewdell WT, Bell JC, McNulty SM, Duda Z, O’Neill RJ, Sullivan BA, Straight AF. RNA-dependent stabilization of SUV39H1 at constitutive heterochromatin. Elife. 2017

Kabeche, L., Nguyen, H. D., Buisson, R. & Zou, L. A mitosis-specific and r loop–driven atr pathway promotes faithful chromosome segregation. Science 359, 108–114 (2018).

Lee, H., Lee, KW., Lee, T. et al. Performance evaluation method for read mapping tool in clinical panel sequencing. Genes Genom 40, 189–197 (2018)

Legros, P.; Malapert, A.; Niinuma, S.; Bernard, P.; Vanoosthuyse, V. RNA processing factors Swd2.2 andSen1 antagonize RNA Pol III-dependent transcription and the localization of condensin at Pol III genes.PLoS Genet.2014

Malig, M., Hartono, S. R., Giafaglione, J. M., Sanz, L. A. & Chedin, F. Ultra-deep coverage single-molecule r-loop footprinting reveals principles of r-loop formation. J. Mol. Biol. (2020).

Miga et al. Telomere-to-telomere assembly of a complete human X chromosome. Nature. 2020

Niehrs C, Luke B. Regulatory R-loops as facilitators of gene expression and genome stability. Nat Rev Mol Cell Biol. 2020

O'Neill RJ. Seq'ing identity and function in a repeat-derived noncoding RNA world. Chromosome Res. 2020 Mar;28(1):111-127. doi: 10.1007/s10577-020-09628-z.

Pan H, Jin M, Ghadiyaram A, Kaur P, Miller HE, Ta HM, Liu M, Fan Y, Mahn C, Gorthi A et al (2020) Cohesin SA1 and SA2 are RNA binding proteins that localize to RNA containing regions on DNA. Nucleic Acids Res

Proudfoot, N. J. Transcriptional termination in mammals: Stopping the rna polymerase ii juggernaut. Science 352 (2016).

Rosenbloom KR et al. (2015) The UCSC Genome Browser database: 2015 update Nucleic Acids Res 43:D670–681 doi:10.1093/nar/gku1177

Roy, D.; Lieber, M.R. G clustering is important for the initiation of transcription-induced R-loopsin vitro,whereas high G density without clustering is sufficient thereafter.Mol. Cell. Biol.2009

Sagie S, Toubiana S, Hartono SR, Katzir H, Tzur-Gilat A, Havazelet S, Francastel C, Velasco G, Chédin F, Selig S. Telomeres in ICF syndrome cells are vulnerable to DNA damage due to elevated DNA:RNA hybrids. Nat Commun. 2017

Sanz, L. A. et al. Prevalent, dynamic, and conserved r-loop structures associate with specific epigenomic signatures in mammals. Mol. cell 63, 167–178 (2016).

Skourti-Stathaki, K. et al. R-loops enhance polycomb repression at a subset of developmental regulator genes. Mol. cell 73, 930–945 (2019).

Sollier, J. & Cimprich, K. A. Breaking bad: R-loops and genome integrity. Trends cell biology 25, 514–522 (2015).

Vanoosthuyse V. Strengths and Weaknesses of the Current Strategies to Map and Characterize R-Loops. Noncoding RNA. 2018 Mar 27;4(2):9. doi: 10.3390/ncrna4020009.

Velazquez Camacho O, Galan C, Swist-Rosowska K, Ching R, Gamalinda M, Karabiber F, De La Rosa-Velazquez I, Engist B, Koschorz B, Shukeir N, Onishi-Seebacher M, van de Nobelen S, Jenuwein T. Major satellite repeat RNA stabilize heterochromatin retention of Suv39h enzymes by RNA-nucleosome association and RNA:DNA hybrid formation

Wahba, L., Costantino, L., Tan, F. J., Zimmer, A. & Koshland, D. S1-drip-seq identifies high expression and polya tracts as major contributors to r-loop formation. Genes & development 30, 1327–1338 (2016).

Wongsurawat T, Jenjaroenpun P, Kwoh CK, Kuznetsov V. Quantitative model of r-loop forming structures reveals a novel level of rna–dna interactome complexity. Nucleic Acids Res. 2012

Xu, W. et al. The r-loop is a common chromatin feature of the arabidopsis genome. Nat. plants 3, 704–714 (2017)

Yan Q, Sarma K. MapR: A Method for Identifying Native R-Loops Genome Wide. Curr Protoc Mol Biol. 2020

Yan, Q., Shields, E. J., Bonasio, R. & Sarma, K. Mapping native r-loops genome-wide using a targeted nuclease approach. Cell reports 29, 1369–1380 (2019).

Yu K, Lieber MR. Current insights into the mechanism of mammalian immunoglobulin class switch recombination. Crit Rev Biochem Mol Biol. 2019 Aug;54(4):333-351. doi: 10.1080/10409238.2019.1659227.

Zeng et al. Association analysis of repetitive elements and R-loop formation across species Mobile DNA (2021) 12:3 https://doi.org/10.1186/s13100-021-00231-5