Alma Mater Studiorum – Università di Bologna

DOTTORATO DI RICERCA IN

Biologia Cellulare e Molecolare

Ciclo XXXII

Settore Concorsuale: 05/E2

Settore Scientifico Disciplinare: **BIO/11**

Multiomic characterization of the effects of G-quadruplex binders on R-loop homeostasis and innate immune response in human cancer cells

Presentata da: Marco Russo

Coordinatore Dottorato

Prof. Giovanni Capranico

Supervisore

Prof. Giovanni Capranico

Esame finale anno 2020

ABSTRACT

R-loops are non-canonical DNA structures consisting of a DNA-RNA hybrid and a displaced ssDNA. R-loops can be structurally compatible with G-quadruplexes, that are secondary DNA structures composed by three stacked tetrads of guanosine. G4s stabilization and R-loops accumulation in cells have been associated with genomic instability and DNA damage. Here, through bioinformatic analysis of genomic R-loop maps and the development of a specific tool to better annotate R-loop regions to gene, we propose that stabilization of G4 structures with specific ligands leads to accumulation of DNA damage and genome instability in cancer cells and that this process is mediated by R-loop stabilization. Moreover, we found that G4 ligand pyridostatin stimulates formation of micronuclei, a hallmark of genome instability. Micronuclei accumulation has been associated with innate immune response triggering through cGAS/STING pathway activation. Activation of this pathway leads to the expression of cytokines, interferons and interferon-stimulated genes. Innate immune system modulation has been proposed as a promising therapeutic strategy for cancer treatment. Here, through analysis of RNA-seq data, we demonstrate that pyridostatin may induce an innate immune stimulation through micronuclei induction in cancer cells. Moreover, to clarify the role of cGAS/STING pathway and the effects of its perturbation in human tumor tissues, we analyzed mutations and expression levels of genes involved in this pathway across 31 cancer types and ~7800 tumor samples from The Cancer Genome Atlas (TCGA). Alterations in mutation status or expression in these genes have been related with innate immune response activation and patient survival and other immune tumor microenvironment features. Our findings indicate that these genes are rarely mutated in human cancers, while their expression may affect the interaction of the tumor with host immune cells affecting disease progression and patient survival.

INDEX

1. INTRODUCTION

1.1 Non-B DNA structures:

The canonical right-handed, double helical form of deoxyribonucleic acid (DNA), known as B-DNA, is the most common DNA secondary structure found in cells. After the discovery of this structure (Watson and Crick, 1953), many other DNA conformations were characterized in a biological relevant context, such as G-quadruplexes (Gellert, Lipsett, & Davies, 1962) and R-loop (Richardson, 1975), highlighting that nucleic acids can form highly dynamic secondary structures, which might have a role in cellular processes.

1.1.1 R-loop

R-loops are composed by a DNA:RNA hybrid duplex and a displaced single DNA strand (Reaban et al., 1994) (Figure 1-1). R-loops are a mainly co-transcriptional event and the mechanism of their formation is explained through the "thread-back" model: during transcription, DNA:RNA hybrid formation is due to the re-annealing of nascent RNA on template DNA. This model is consistent with crystallographic and biochemical evidence showing that RNA polymerases actively separate the template DNA strand from the nascent RNA, which exit from the RNA polymerase through two distinct channels (Westover, 2004).

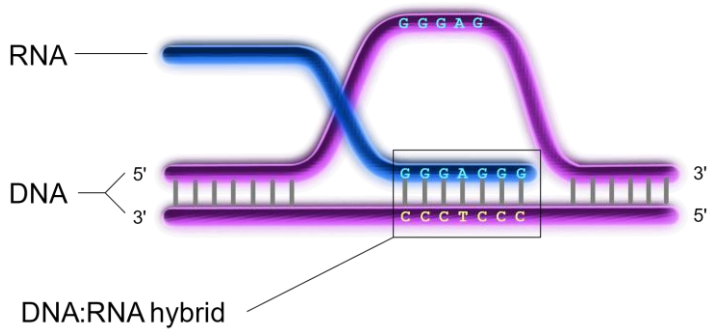

Figure 1-1 Schematic representation of R-loop structure. Image licensed under the Creative Commons Attribution 3.0 Unported license.

Over the years, many studies explained that R-loop structures forms with different efficiency depending on:

a. DNA topology: one of the major favouring event of R-loop formation is the negative supercoiling of DNA that occurs behind the elongating RNA polymerase (Drolet et al., 1994). In fact, DNA topoisomerase activity, which resolves the negative supercoiling of DNA, can prevent R-loop formation as demonstrated in vitro and in E. coli (Phoenix et al., 1997). Recent studies in human cancer and yeast cells highlight how eukaryotic DNA topoisomerase I can regulate R-loop dynamics depending on the genomic context of R-loop formation(El Hage et al., 2014; Manzo et al., 2018).
b. Sequence composition: it has been reported in *in vitro* studies how the presence of guanosine clusters in the starting region of R-loop results in a highly efficient production of this structure (Roy and Lieber, 2009). This is due to the fact that G-rich-RNA:DNA hybrids are more stable than DNA:DNA duplexes with the same sequence (Ratmeyer et al., 1994). More recently, it was reported that CpG islands, which are the most common mammalian class II promoters and can have an asymmetry in the distribution of guanine and cytosines between the two strands (property known as “GC-skew”), are more prone to form R-loop (Ginno et al., 2012).

These findings are consistent with the fact that R-loop structures form prevalently at transcription start and termination sites of active genes, characterized by high GC-skew (Ginno et al., 2013).

Several factors can regulate R-loops, in particular helicases, such as Bloom, Senataxin and Acquarius, are known to resolve the hybrid duplex of R-loops (Chang et al., 2017; Skourti-Stathaki et al., 2011; Sollier et al., 2014). In addition, R-loop hybrid structure can be degraded by the action of RNAseH enzymes (Wahba et al., 2011), which are nucleases specific for RNAs annealed to a DNA strand. In human, two such RNases exist: RNaseH1 and RNaseH2, which can both resolve R-loop hybrids (Wahba et al., 2011). Unscheduled R-loop formation can indeed be rescued by overexpressing *RNAseH1* in mammalian cells (Domínguez-Sánchez et al., 2011).

1.1.1.1 R-loops as transcriptional regulator

While for many years R-loops were considered by-products of the transcription process, many evidences of their involvement in different cellular processes have been reported. Since gene promoters and terminators are a hotspot of R-loop formation, their involvement in transcriptional regulation was then easily hypothesized. In fact, R-loops can favour transcription through protecting promoters from methylation (Grunseich et al., 2018) or preventing transcriptional repressor binding (Chen et al., 2015). Furthermore, a role for R-loops in transcription termination has been demonstrated (Skourti-Stathaki et al., 2011). R-loop formation in G-rich regions in proximity of the poly(A) signal causes RNA polymerase II pausing, whereas R-loop resolution by senataxin and RNA degradation by *Xrn2* exonuclease allow RNAPolII transcription termination. Moreover, at transcription terminator sites of some genes, R-loop formation can induce repressive chromatin remodelling that favours RNA PolII pausing and transcription termination (Skourti-Stathaki et al., 2014).

1.1.1.2 R-loops as genomic threat

While in determined circumstances R-loops have an important role in mediating physiological processes, on the other hand their unscheduled formation can be a source of DNA damage and genome instability. One of the main example of mechanisms of DNA damage in which R-loops are involved is the "transcription-replication conflict", which occurs when the transcription machinery collides with a replication fork. In fact, it has been demonstrated, in bacteria, in yeast and in human cell lines, that R-loop mediated genome instability can occur during S phase (Gan et al., 2011; Wellinger et al., 2006). Moreover, it was observed that the severity of damage caused by transcription replication conflicts depends on their orientation: while "head-on" conflicts favour R-loop formation and lead to more severe genomic instability, co-directional conflicts are resolved by replicative helicases that remove R-loop structures (Hamperl et al., 2017) (Figure 1-2). These two types of conflicts can specifically activate DNA damage responses mediated by either ATM or ATR kinases.

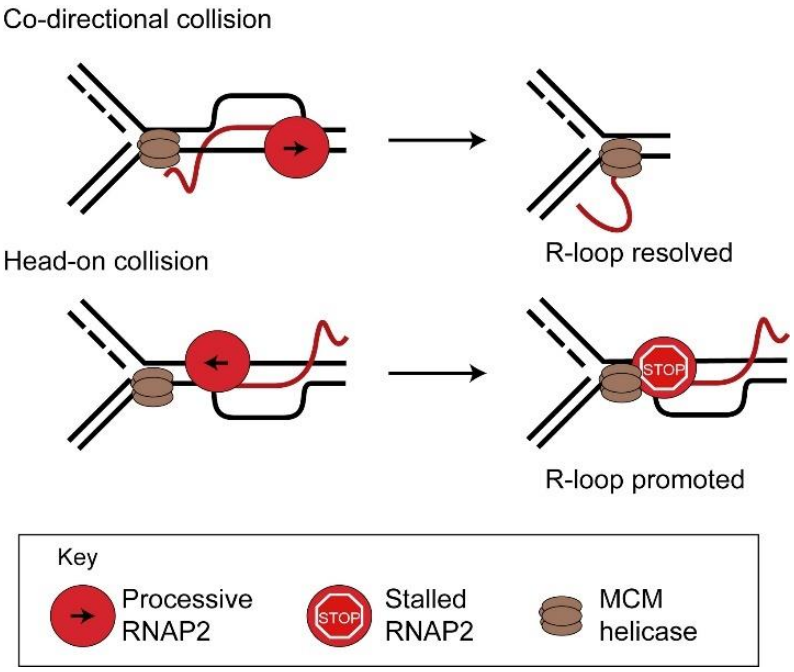

Figure 1-2 Schematic representation of R-Loop involvement in Transcription-Replication Conflicts. Adapted from Crossley et al., 2019*. Image licensed by Elsevier and Copyright Clearance Center (License n. 4698220460707).*

While, as reported previously, R-loop presence may favour transcription event in particular promoter condition, it has also been reported that in particular condition R-loop may repress transcription (Bonnet et al., 2017; Hamperl et al., 2017).

Unscheduled formation of R-loop has been associated with a large number of pathologies. Often, the loss of genes the function of which helps to prevent or resolve R-loop structures, like *BRCA2* (Shivji et al., 2018), senataxin (Becherel et al., 2015), *TDP1* (Cristini et al., 2019), *TREX1* and *RNaseH2* (Lim et al., 2015), TDRD3-TOP3B complex (Yang et al., 2014) causes different diseases such as cancer, neurodegeneration and immune related syndromes.

1.1.1.3 R-loop detection

Genomic R-loop detection is usually obtained through S9.6 antibody that specifically recognizes DNA:RNA hybrids (Boguslawski et al., 1986). This antibody can be used for both immunofluorescence microscopy and immunoprecipitation, a technique known as DNA:RNA

ImmunoPrecipitation or DRIP (Ginno et al., 2012). One of the first genome-wide R-loop detection method (DRIP-seq), which is still broadly used, is based on the DRIP protocol followed by next generation sequencing (Ginno et al., 2012). In recent years, many variant techniques were derived from the original DRIP to map genomic R-loops. Most of these variants have improved the first protocol, such as S1-DRIP-seq (Wahba et al., 2016), RDIP-seq (Nadel et al., 2015), ssDRIP-seq (Xu et al., 2017). Other methods are based on potassium permanganate footprinting (Kouzine et al., 2017) cytosine deamination by bisulphite or a catalytically inactive RNAseH to recognize the RNA:DNA hybrid.

The most relevant techniques are:

- DRIP-seq (Ginno et al., 2012): as already said, this technique uses S9.6 antibody to immunoprecipitate DNA:RNA hybrids regions. Before immunoprecipitation genomic DNA is gently extracted from cells, to avoid, R-loop disruption and fragmented through restriction enzyme digestion. After immunoprecipitation recovered DNA is sequenced (Figure 1-3). The principal disadvantages of DRIP-seq are the lack of R-loop strandness information and the low resolution (usually kilobases) of R-loop signal compared to other techniques. A close variant of the DRIP protocol, which includes a sonication step of genomic DNA instead of a restriction enzyme digestion (Halász et al., 2017), has an improved resolution of R loop peak.

- DRIPc-seq (Sanz et al., 2016): this technique is an improvement of the DRIP-seq protocol in which after DNA:RNA immunoprecipitation, RNA strand is recovered, retrotranscribed and sequenced. Although the protocol is more complex compared to DRIP-seq protocol, this method provides a higher resolution signal, and the strandness information of the R-loop.
- Bis-DRIP-seq (Dumelie and Jaffrey, 2017): bisulphite-DNA-RNA immunoprecipitation sequencing is another improvement of DRIP-seq protocol in which bisulphite is used to convert cytosine to uracil residues in single strand DNA of the R-loop. This allow to recognize R-loops region with strand specificity.

- DRIVE-seq (Ginno et al., 2012): DNA:RNA in vitro enrichment (DRIVE) sequencing was the first method that does not rely on S9.6 capture of R-loop. This technique uses a catalytically inactive RNASEH1 that binds to DNA:RNA hybrid structure to perform an in vitro pull-down assay. This method can be useful as to validate S9.6-captured R-loops, since it is based on a different isolation approach.
- R-ChIP-seq (Chen et al., 2017): This method uses the catalytically inactive RNASEH1 to recognize R-loops, but differently from DRIVE-seq, this enzyme is expressed in cells to allow the in vivo capture of R-loop and then a chromatin immunoprecipitation is performed.

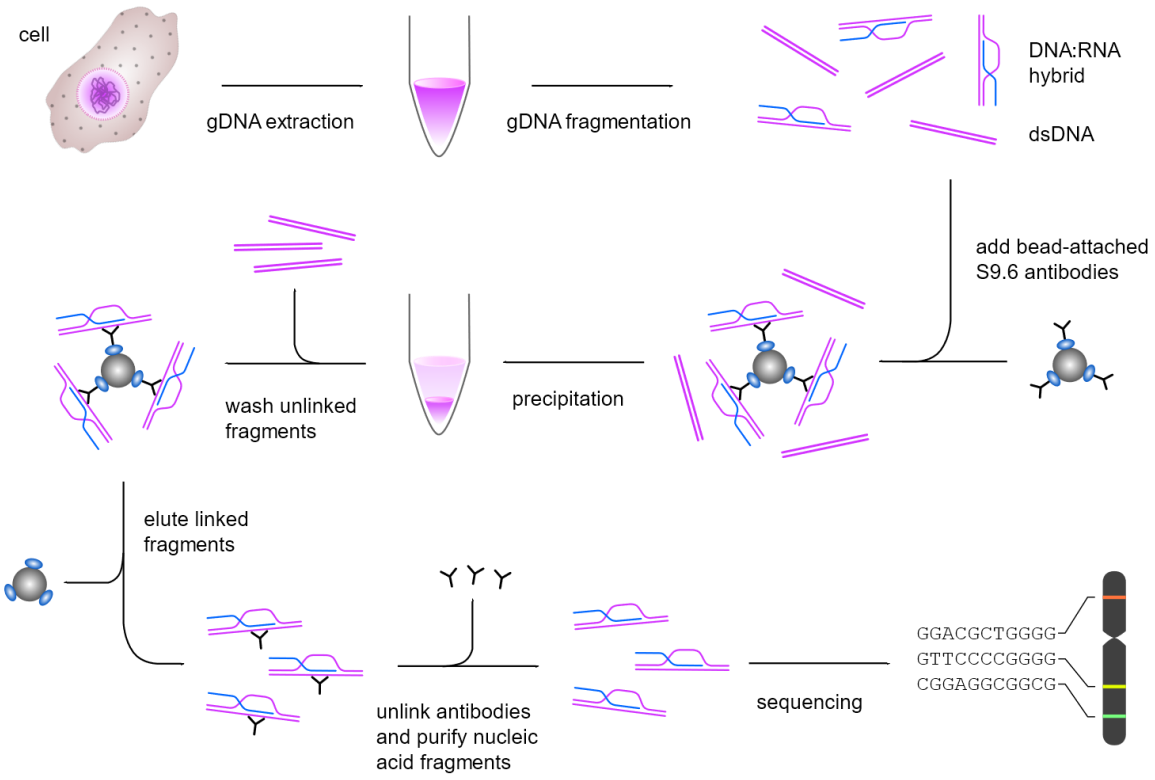

Figure 1-3 Workflow of DRIP (DNA:RNA Immunoprecipitation) sequencing experiment. Image licensed under the Creative Commons Attribution-Share Alike 3.0 Unported license.

Despite of the variety of techniques to map R-loop over the years, DRIP-seq remains the most used in studies that involves genomic R-loop detection, due to the easier set up compared to other methods.

1.1.2 G-quadruplexes

Since R-loop formation is favoured by guanosine-rich regions, it was demonstrated in vitro that displaced DNA strands of R-loop can harbour G quadruplex structures (Duquette et al., 2004).

G-quadruplexes (G4s) are secondary DNA structure composed by three stacked tetrads of guanosines (Gellert et al., 1962) (Figure 1-4). While the most common G4 structure is composed by a single strand of DNA (intramolecular G4), they can also be formed by more than one strand of DNA (intermolecular) and even by an RNA and a DNA strands at the same time (Xiao et al., 2013).

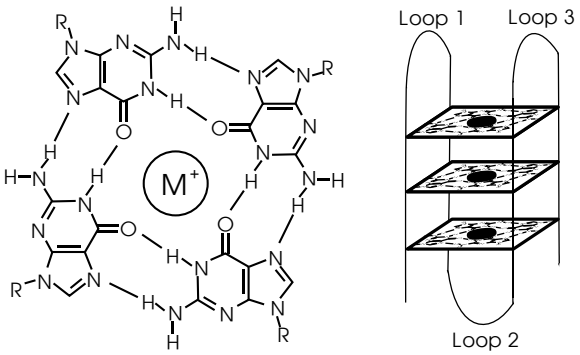

Figure 1-4 Schematic representation of a guanosine tetrads (left) and of a intramolecular G-quadruplex composed of three stacked tetrads (right). Image licensed under the Creative Commons Attribution-Share Alike 2.5 Generic license.

1.1.2.1 G-quadruplex detection

As G4 structures are highly polymorphic, depending mainly on the strand involved and the size of sequence loops between two quartet guanines (Kwok and Merrick, 2017), the prediction of G4 formation and stability in a given sequence is not straightforward and has been proven difficult. Nevertheless, based on a defined guanosine motif sequence, genomic regions with the potential to form G4s can be identified using in silico motif search algorithms. The defined generic G4 motif is formed by four stretches of 3-5 guanines separated by 1-7 nucleotides:

$$G_{(3-5)}\text{-}N_{(1-7)}\text{-}G_{(3-5)}\text{-}N_{(1-7)}\text{-}G_{(3-5)}\text{-}N_{(1-7)}\text{-}G_{(3-5)}$$

During the last years, other more refined sequence detection tool were developed, based on Markov window model, such as Quadparser (Huppert and Balasubramanian, 2005), or on other parameters like G-richness and G-skewness of the sequence, such as G4hunter (Bedrat et al., 2016). While the number of G4 promoting sequences on the human genome can vary on the base of the tool used, all the tools confirmed that G4s are not randomly distributed in mammalian genomes, and are particularly enriched at telomeres, promoters, transcription terminations and replication origins.

To confirm all the *in-silico* findings, different methods were developed to recognize G4 structures in vivo and genome-wide. In particular a specific antibody, called BG4, was developed (Biffi et al., 2013), which has been successfully used both for visualization of G4s by immunofluorescence microscopy and for determination of G4s along the genome by chromatin immunoprecipitation technique and next generation sequencing (Hänsel-Hertsch et al., 2016; Mao et al., 2018) (Figure 1-5). G4-ChIP data on human cell lines have allowed to get more insight on G4 localization along the genome, and suggested that G4s are localized in many regulatory elements, like highly expressed gene promoters, such us the proto-oncogene *MYC* (Hänsel-Hertsch et al., 2016; Sun and Hurley, 2009).

Moreover, another genome-wide approach, based on the polymerase stop assay (a technique based on the inhibition of DNA polymerase activity by a stable secondary structure of the DNA template) and next-generation sequencing was developed and used to map G4

regions in genomic DNA *in vitro* (Chambers et al., 2015) (Figure 1-5). This technique allowed to identify G4s formed in human genome stabilized by either a cation (K+) or pyridostatin (a specific ligand of G4s). Data produced with this methodology, which has been further improved recently (Marsico et al., 2019), lead to the development of Quadron, a tool based on machine learning approach, that on the base of known G4 structure sequences, can predict G4 presence in any other DNA genomic sequence (Sahakyan et al., 2017). Combination of G4-seq techniques and prediction tools allowed to determine G4 localization on genomic DNA of 12 different species (Marsico et al., 2019).

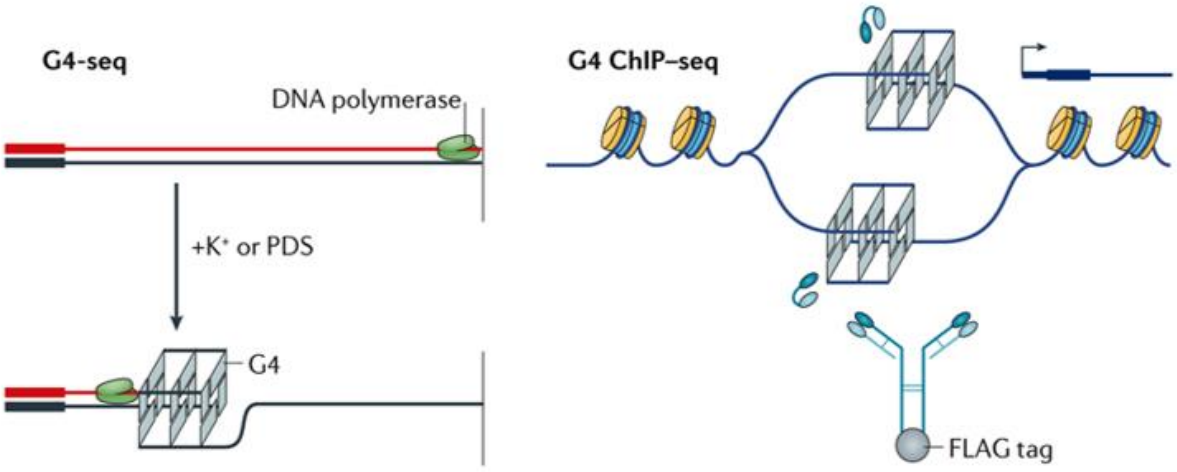

Figure 1-5 Schematic principle of G4-seq (left) based on DNA polymerase stalling at G4s, and G4 ChIP-seq (right), that takes advantage of BG4 antibody to recognize G4 structures. Adapted from Hänsel-Hertsch et al., 2017. *Image licensed by Springer Nature and Copyright Clearance Center (License n. 4698220569761).*

Thanks to these efforts, more than 700,000 G4 promoting regions were identified in the human genome. The determination of genomic localization of G4 structures allows to further investigate their interplay with other non-canonical structures (such as R-loops) and to characterize their role in genomic regulatory regions.

1.1.2.2 G-quadruplexes function

In recent times, a growing interest has emerged on G4 dynamics and their involvement in cellular processes such as transcription and DNA replication. G4s were reported to be enriched upstream of many gene transcription start site and in nuclease hypersensitive regions (Huppert and Balasubramanian, 2007), suggesting a potential role in transcription regulation. Moreover, they have been associated with two transcription-associated helicases (XBP and XPD), the ability of which to bind and resolve G4s has been demonstrated (Gray et al., 2014). While G4s were reported to efficiently inhibit transcription at some loci (Figure 1-6), like in c-Myc promoter (Siddiqui-Jain et al., 2002), more recently G4s were observed at highly transcribed gene promoters and their stabilization have been associated with increased transcriptional activity (Hänsel-Hertsch et al., 2016). Furthermore, their presence was correlated with CpG island hypomethylation through inhibition of DNMT1 (a DNA methyltransferase) activity (Mao et al., 2018) (Figure 1-6).

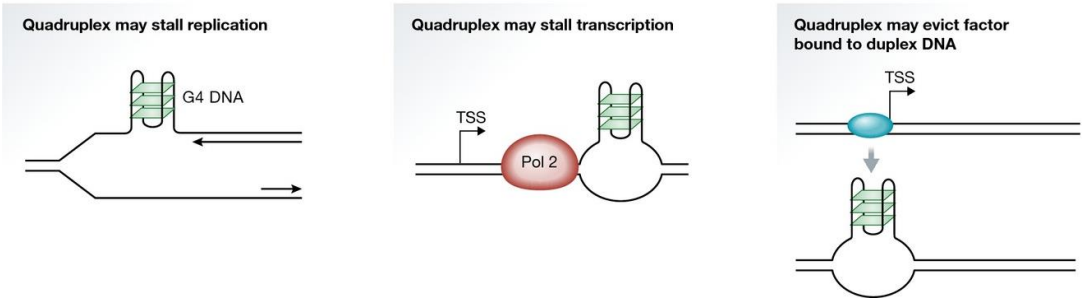

Figure 1-6 schematic representation of G-quadruplex effects during replication and transcription. G4s are involved in polymerase stalling in replication forks. During transcription, G4s may have different roles depending on their localization. They can cause RNA polymerase stalling, leading to transcription block, or protect DNA from factors that methylate DNA, favouring gene expression. Adapted from (Maizels, 2015). *Image licensed by John Wiley and Sons and Copyright Clearance Center (License n. 4698220652919).*

Another process in which G4s are involved is DNA replication. As mentioned above, DNA polymerase stall at G4 structures, that can then cause replication arrest (Figure 1-6). Many helicases can resolve these structures to ensure proper replication fork progression. Two of

these, BLM and WRN, are involved in telomere maintenance. Mutations of these genes (which cause Bloom syndrome and Werner syndrome, respectively) lead to G4 accumulation at telomeres and genome instability (Crabbe et al., 2004; Drosopoulos et al., 2015; Sun et al., 1998). Other helicases operate at internal genomic region, such as FANCJ, which mutations can cause chromosomal instability, Fanconi anaemia and some breast and ovarian cancer types (Brosh and Cantor, 2014).

1.1.2.3 G-quadruplexes binders

In the last 30 years, around 1000 different molecules that recognize and stabilize G4s were developed. Many of them were initially developed to target G4 structures at telomeres. Among all these compounds, one of the most used G4 binder is pyridostatin, or PDS (Rodriguez et al., 2008) (Figure 1-7). Reported effects of PDS treatment in cancer cell lines are DNA cleavage accumulation, cell cycle arrest and activation of DNA damage response (Rodriguez et al., 2012). Interestingly, long time treatment induces PARP1 cleavage and apoptosis of cancer cells. Moreover, in cancer cells lacking BRCA1, BRCA2 or RAD51 (involved in homologous recombination repair) researchers have observed an increased G4 formation, which is exacerbated by PDS suggesting its potential role in cancer therapy as a cytotoxic compound (Zimmer et al., 2016).

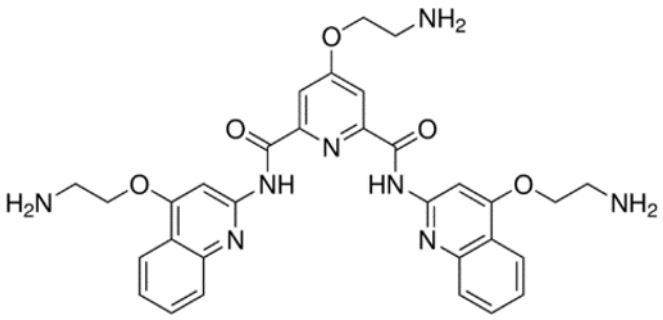

Figure 1-7 Pyridostatin (4-(2-Aminoethoxy)-N2,N6-bis[4-(2-aminoethoxy)-2-quinolinyl]-2,6-pyridine-dicarboxamide) chemical structure.

Another selective compound developed to specifically bind G4s is BRACO-19, designed to efficiently target G4s at telomeres and inhibit telomerase activity. Braco-19 derivatives have been proposed as potential anti-telomere agents in therapeutic approaches against human cancers (Read et al., 2001) (Figure 1-8).

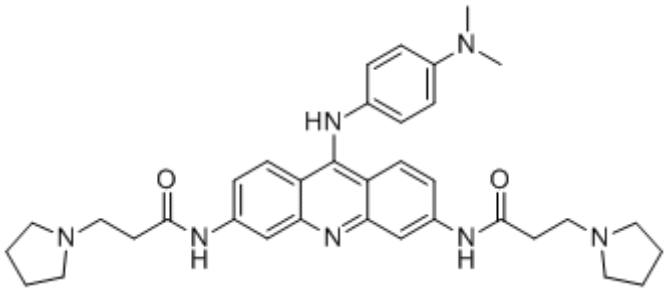

Figure 1-8 BRACO-19 (N,N'-(9-(4-(Dimethylamino)phenylamino)acridine-3,6-diyl)bis(3-(pyrrolidin-1-yl) propanamide)) chemical structure.

At the University of Bologna, another G4 stabilizer, called FG (Figure 1-9), was recently developed along with several other analogues in a hydrazine series, which have been shown to specifically recognize and stabilise G4s in different human cancer cell lines (Amato et al., 2016), suggesting a potential role in cancer therapy, as previously described compounds.

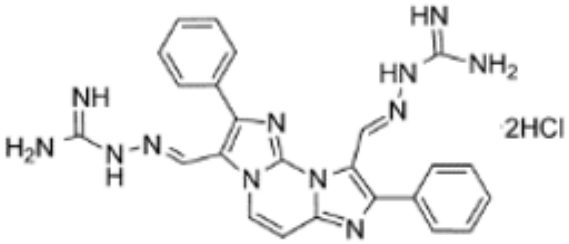

Figure 1-9 FG ((2E,2'E)-2,2'-((2,8-diphenyldiimidazo[1,2-a:1',2'-c]pyrimidine-3,9-diyl)bis(methanylylidene))bis(hydrazine-1-carboximidamide)) chemical structure.

1.2 Innate immune response in cancer

Recently, the development of innovative therapeutic strategies for cancer treatment focused on promoting a systemic immune response against the tumour. Results of this emerging field lead to the introduction of clinically approved or potential "T-cell focused" therapies, like PD-1 targeting with antibodies that enhance T-cell response (Okazaki and Honjo, 2007), anticancer vaccines that recognizes neoantigens expressed by cancer cells (Hollingsworth and Jansen, 2019) or infusion of engineered T-cells with chimeric antigen receptors (CAR T-cells) to specifically recognize cancer cells (Kalos et al., 2011).

However, successfully T-cells activation required a previous induction of innate immune response processes (Medzhitov and Janeway, 1999). Tumour cells recognition by the innate immune response machinery does not rely on anti-cancer specific sensors but makes use of pathways that are also used to recognize external pathogens, like bacteria and viruses. One of the main pathways involved in innate immunity, which is based on sensing cytoplasmatic nucleic acids, is the cGAS/STING pathway.

1.2.1 cGAS/STING pathway

The cGAS/STING pathway main actors are "Cyclic GMP-AMP synthase" (*cGAS*) (Sun et al., 2013) and "Stimulator of Interferon Genes"(*STING*) genes. cGAS is a cytosolic DNA sensor, which catalyses the synthesis of Cyclic guanosine monophosphate–adenosine monophosphate (cGAMP) upon binding to cytosolic DNA. cGAMP is recognized and bound by STING, a transmembrane protein localized in the endoplasmic reticulum. This binding leads to STING activation and phosphorylation of TANK-binding kinase 1 (TBK1), which in turn phosphorylates Interferon regulatory factor 3 (IRF3), a transcription factor that after activation translocates into the nucleus inducing transcription of Type I interferons gene (Figure 1-10).

The cGAS/STING pathway is required in immune response against tumours. It has been reported that this pathway is required to obtain T-cell recruitment in tumour environment

and that it can be activated by self-DNA derived by tumour cells (Woo et al., 2014). For this reason, boosting of this pathway is considered an interesting therapeutic approach as a coadjutant in cancer therapies.

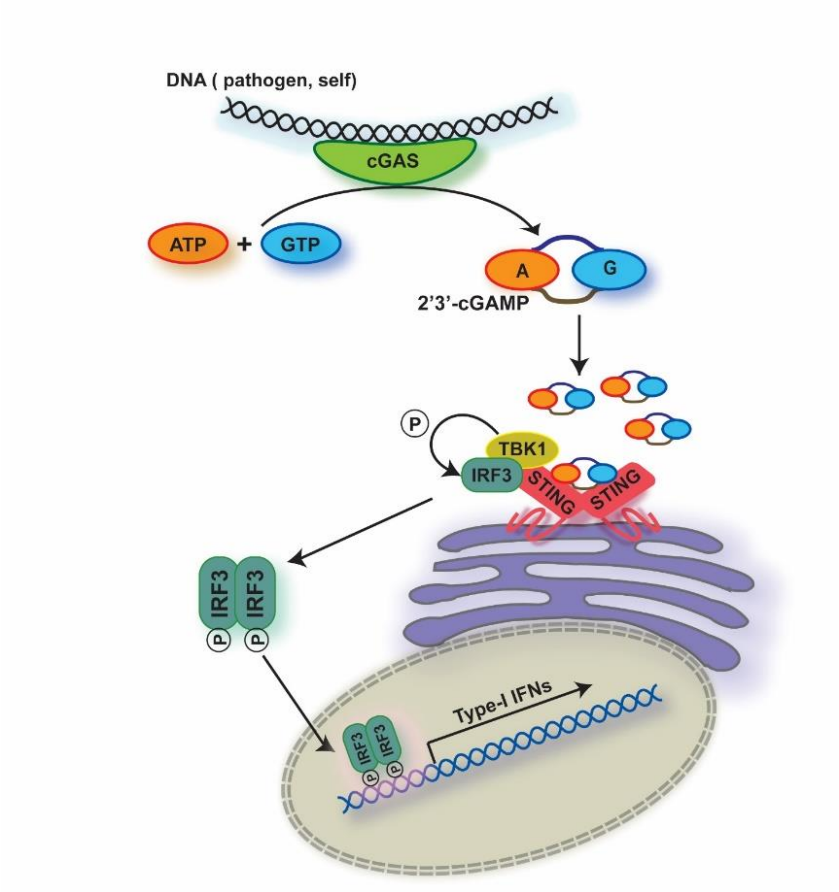

Figure 1-10 Schematic representation of cGAS/STING pathway activation. Image licensed under the Creative Commons Attribution-Share Alike 3.0 Unported license.

1.2.2 Innate immune activation can be mediated by micronuclei

Self-DNA is often present in cancer cell cytosol because of micronuclei formation. Micronuclei are chromatin fragments surrounded by a complete nuclear double-layer membrane that are generated because of a failure in chromosome segregation and cell cytokinesis at mitosis. Micronuclei formation in normal cells has been used as a marker of genotoxicity and genome instability caused by pollutants and toxin chemicals (Shelby, 1988).

Once micronuclei are formed, they can be removed by the autophagy pathway (Bartsch et al., 2017; Lan et al., 2014) or undergo disruption in cytoplasm (Hatch et al., 2013). Moreover, it was reported that micronuclei are not only a “side effect” of genome instability, but can also be trigger of further and peculiar genomic rearrangements of entire or large portions of single chromosomes, a phenomenon known as chromothripsis (Zhang et al., 2015).

Recently, it has been reported that micronuclei can induce an innate immune response via the cGAS/STING pathway through binding of cGAS to micronuclear DNA (MacKenzie et al., 2017). In another study (Harding et al., 2017), authors highlighted that inflammatory response mediated by radiation-induced micronuclei leads to a systemic response against the tumour through a paracrine signalling mechanism. Moreover, cGAS/STING activation mediated by micronuclei formation has been observed also in cells with *BRCA2* deficiency (Heijink et al., 2019) and in *BLM* helicase deficient cells (Gratia et al., 2019). These findings suggest that several types of DNA damage, may favour micronuclei formation and consequently innate immune response activation.

However, it remains to be established whether DNA-damaging small molecules may promote an increase of cytosolic DNAs and may efficiently activate an innate immune response in human cancers. Such agents might then be used in clinical treatments of cancer patients as immunomodulators rather than cell killing drugs.

1.3 Aim of the project

The first aim of the present PhD thesis was to understand, through bioinformatic analysis, how R-loop homeostasis changes in human cancer cells in presence of G4-binding compounds to characterize the mechanism of increased R-loop by G4-binders. In particular, using DRIP-sequencing technique we produced R-loop maps in U2OS cancer cells treated with PDS, FG and BRACO-19 and in control conditions to study how R-loops alterations correlate with G4 position in human genome.

To perform this analysis, we have first addressed the problem of lack of strand information (required to better characterize the structural interplay between R-loops and G4s) using DRIP-seq data. To do so, we developed a genome annotation tool, called DRIP-seq Optimized Peak Annotator (DROPA), that we used to predict R-loop strandness.

The second aim was to establish whether G4 binders can activate an innate immune response in human cancer cells. To this end, we determined the transcriptome profiles in G4 binder-treated and untreated cells by preparing RNA-seq libraries of human breast cancer MCF-7 cells.

Lastly, in collaboration with Prof. Francesca Ciccarelli (The Francis Crick Institute, London, UK), in whose laboratory I spent four months as a visiting PhD student, we investigated publicly available genomic data (from The Cancer Genome Atlas) of human cancer and normal tissues to find cancer types showing significant mutations in genes of the innate immune response, in particular regarding the cGAS/STING pathway.

2. METHODS

2.1 DRIP-sequencing data analysis

We used DNA:RNA immunoprecipitation (DRIP) methodology to immunoprecipitate and isolate DNA:RNA duplexes from genomic DNA preparations by using S9.6 antibody and to map genome-wide R-loop structures. Full DRIP protocol is described in De Magis et al., 2019.

DRIP immunoprecipitates obtained from 40 micrograms of digested genomic DNA were pooled together and sonicated. Ligation of Illumina Truseq adapters was performed according to manufacturer's instructions. All DRIP-sequencing libraries produced by our laboratory were sequenced by Illumina HiSeq4000 platform (pair-end 2X150 bp) at Biodiversa S.r.l. (Rovereto,TN, Italy)

2.1.1 DRIP-seq peak calling

All the DRIP-seq libraries followed a standard peak calling workflow (Figure 2-1).

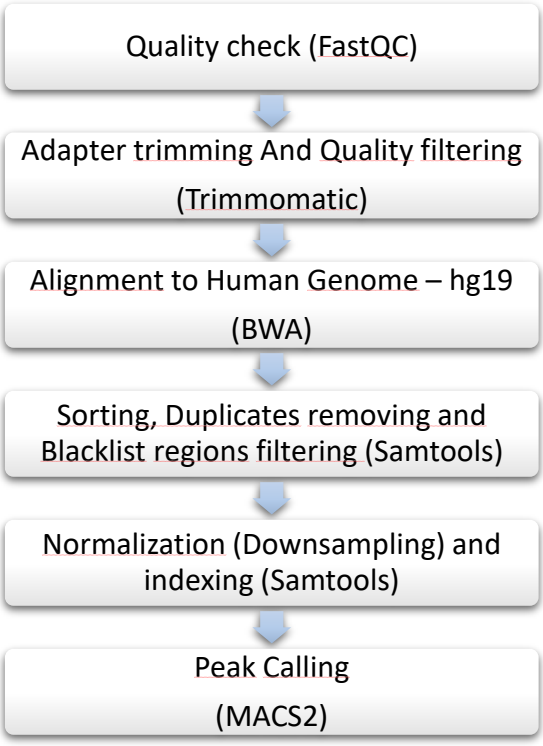

Figure 2-1 Workflow of DRIP-seq peak calling.

2.1.1.1 FastQC

FastQC (version 0.11.8) is a tool that provide some quality control checks with summary graphs and tables on raw sequence data coming from high throughput sequencing platforms. The command line used for each sequencing library was in the format:

```
> fastqc -o fastqc/folder -t 2 input.fastq
```

Input.fastq is the fastq file. Quality control of reads is performed before and after trimming step.

2.1.1.2 Trimmomatic

Trimmomatic (version 0.33) is a tool that allow to remove adapters and low quality bases from sequenced reads (Bolger et al., 2014). The command line used for each pair of sequencing library was in the format:

```
> java -jar trimmomatic-0.33.jar PE input_fwd.fq.gz input_rev.fq.gz
output_fwd_paired.fq.gz output_fwd_unpaired.fq.gz
output_rev_paired.fq.gz output_rev_unpaired.fq.gz
ILLUMINACLIP:TruSeq3-PE.fa:2:30:10:2 LEADING:20 TRAILING:20 MINLEN:25
```

Where “input_fwd.fq.gz” and “input_rev.fq.gz” are the paired end input fastq files, “output_fwd_paired.fq.gz” and “output_rev _paired.fq.gz” are the paired end fastq files that passed the trimming process and “output_fwd_unpaired.fq.gz” and “output_rev_unpaired.fq.gz” are the reads that were discarded because of low quality score. “ILLUMINACLIP:TruSeq3-PE.fa:2:30:10:2” arguments regards the type of adapter that will be removed, “LEADING:20” and “TRAILING:20” arguments will allow to remove leading or trailing bases of each reads if they have a quality value below 20. “MINLEN:25” argument will remove every read that is below 25 bases long.

2.1.1.3 BWA

Burrow-Wheeler Aligner (version 0.7.13) is a software that allows to map reads against a reference genome, such as the human genome. The genome reference used was the human genome hg19 version. The command line used for each pair of sequencing library was in the format:

```
> bwa aln -n 2 -t 8 index.hg19 output_fwd_paired.fq.gz > for.sai
> bwa aln -n 2 -t 8 index.hg19 output_fwd_paired.fq.gz > rev.sai
> bwa sampe $IND for.sai rev.sai output_fwd_paired.fq.gz >
output_fwd_paired.fq.gz > output.bwa.sam
```

The first two commands align reads on human genome, while the last one creates the sequence alignment map (SAM) file.

2.1.1.4 SAMtools

Samtools (version 1.8) is a suite of tool for manipulating SAM file (Li et al., 2009). The command line used for SAM file was in the format:

```
> samtools view -b -h -o output.bwa.bam output.bwa.sam
> samtools sort -o output.bwa.bam -@ 8 output.sorted.bam
> samtools rmdup output.sorted.bam output.rmdup.bam
> samtools index output.rmdup.bam
```

The command *view* was used to convert SAM files to BAM, filter blacklist regions, and downsample libraries to the smallest one. The command *sort* was used to sort alignment. The command *rmdup* was used to remove duplicated reads. The command *index* was used to index BAM files.

2.1.1.5 MACS2

MACS2 (version 2.1.0) is a software designed to identify genomic regions (called peaks) in which there is a significative enrichment of read (Zhang et al., 2008). The command line used for SAM file was in the format:

```
> macs2 callpeak -t output.rmdup.bam -c input.bam -f BAMPE --keep-dup
all -g hs --outdir Sample -n Sample -B
```

For each library we used a fragmented genomic DNA library as control, and we obtain a peak file in BED format.

2.1.2 DRIP-seq peak processing

DRIP seq libraries regrading pilot experiment (see section 3.1.1) were analysed with ODIN differential peak caller and PAVIS peak annotator.

DRIP-seq libraries and peak data in section 3.1.2, since the presence of biological replicates and the development of optimized annotation tool, were analysed differently. Analysis involved the use of different tools to perform calculation of correlation between replicates, to assess peak consensus between replicates, to visualize data and to perform differential analysis of peak intensity or length. Firstly, tools used will be introduce. Last paragraph will illustrate how analysis was performed.

2.1.2.1 ODIN

ODIN (version 0.3.2) (Allhoff et al., 2014) is a Hidden Markov Model-based tool to call and analyse differential binding of peaks in pairs of ChIP-seq data without replicates. ODIN performs signal processing, peak calling and enrichment calculation. Standard command-line was used:

```
> rgt-ODIN -v -f 1.5 Control_library.bam Treated_library.bam
~/rgtdata/hg19/genome.fa ~/rgtdata/hg19/chrom.sizes --input-
1=INPUT.bam --input-2=INPUT.bam --output-dir=outputodin
```

2.1.2.2 PAVIS

PAVIS (Huang et al., 2013) PAVIS is a web-based tool for peak annotation and visualization of ChIP-seq data. Furthermore, the annotation function reports relative enrichment levels of peaks in different genomic regions. It works with BED files.

2.1.2.3 Deeptools

Deeptools (version 3.0.1) is a suite of commands to manipulate and process BAM files (Ramírez et al., 2016). We used commands multiBamSummary and plotCorrelation to assess correlation between replicates of DRIP-seq libraries with the following standard command line:

```
> multiBamSummary bins -b bamfiles.list -out correlationn.npz -p 8 --
centerReads
> plotCorrelation -in correlationn.npz -o correlationn.svg -p
scatterplot
```

We further used BamCoverage command to create genomic signal profiles in BigWig format, with this command line:

```
> BamCoverage -b output.rmdup.bam -o sample.bw -p 8
```

2.1.2.4 Bedtools

Bedtools (version 2.28.0) is a suite of tools used to manipulate BED files (e.g. output files of MACS2) (Quinlan and Hall, 2010). Bedtools commands that we used in our analysis were:

- Bedtools *intersect* was used to compute the overlap of 2 sets of peaks with the standard command line:

  ```
  > bedtools intersect -a peakfile1.bed -b peakfile2.bed
  ```

- Bedtools *merge* was used to combine genomic regions into a unique one with the standard command line:

  ```
  > bedtools merge -i bedfile.list
  ```

- Bedtools coverage was used to compute intensity of sequencing signal in specific genomic regions using the command line:

  ```
  > bedtools coverage -a signal.bam -b regions.bed
  ```

- Bedtools *shuffle* was used to randomize peaks all over the genome or in specific genomic contexts to perform enrichment analysis. It was used with standard command line:

  ```
  > Bedtools shuffle -i region.bed -g genome.hg19
  ```

  where -g is the genome of interest.

2.1.2.5 Integrative Genomics Viewer (IGV)

Integrative Genomics Viewer (IGV) (version 2.0.1) is a visualization software for interactive exploration of la genomic libraries. It supports a wide variety of data types, including array-based and next-generation sequence data, and genomic annotations (Robinson et al., 2011).

2.1.2.6 SkewR

SkewR is a tool that uses Hidden Markov Model (HMM) to identify GC skew regions in any genome based on trained StochHMM models (Ginno et al., 2012). We used it on human genome hg19 with the command line:

```
> Perl RunGC-Skew.pl -s hg19.fa -m GC_SKEW_7600.hmm -g
hg19_genereference.bed -b hg19_cpg.bed -o Result
```

2.1.2.7 EdgeR

EdgeR (version 3.26.8) is an R/Bioconductor package used to perform differential analysis with different statistical methodologies. This tool was used to perform differential analysis of DRIP-seq peak intensity and peak length between control and treatment condition.

2.1.2.8 DRIP-peak data processing and differential analysis

Analysis of correlation between DRIP-seq experiments replicates and the Pearson correlation coefficient computation was performed with Deeptools suite using multiBamSummary and plotCorrelation commands. Genomic signal tracks were built with Deeptools suite using bamCoverage command and visualized with Integrative Genomics Viewer.

Peak consensus between DRIP-seq replicates was calculated using Bedtools intersect and merge commands.

To annotate DRIP-seq peaks and to establish peaks strand we developed and used DROPA tool (see section 2.2 and 3.1.2).

Signal plot over TSS were calculated using Bedtools coverage command. TSS regions were categorized in four different groups on the base of the gene expression level (top 10%, high expression; mid 57%, intermediate expression; low 33%, low expression; silent, no expression) as established by RNA-seq experiments (high expression, FPKM>57.5; intermediate expression, 5.5<FPKM<57.5; low expression, 1<FPKM<5.5; no expression, FPKM<1). Moreover, TSS groups were also categorized basing on the presence of a CpG island within 4kb window and on the presence of GC skew in CpG island region. GC skew region were defined using SkewR tool with High threshold model set.

Differential analysis of peak signal level was performed using Bedtools coverage command to determine peak read coverage and EdgeR. differential expressed peaks were selected with a p-value < 0.001 and a treated/control fold change > 1.5.

Differential analysis of peak length was performed selecting only peaks that were present both in control and treatment samples with bedtools intersect command. Then a differential length analysis was performed, and peaks were selected only with a treated/control size fold-change > 1.5 and a p-value < 0.05.

To assess enrichment in G4 structures, each extended peak set was randomized, using bedtools shuffle command, on genomic regions containing expressed genes with a CpG island localized within 5 kb from Transcription start site. Matched unchanged peak sets were prepared choosing, for every extended peak, an unchanged peak with the same dimension and the same sub-genic localization (Upstream, Gene body, Downstream) using custom scripts. Randomization was performed in the same way of the extended peak sets. Intersection with G4 set were performed using bedtools intersect command. Statistical significance was determined with the Kolmogorov-Smirnov test.

2.2 DROPA development and performance evaluation

DROPA software was developed in Pyhon3 language and can be launched in UNIX environment from command-line. It is available at https://github.com/marcrusso/DROPA

2.2.1 DROPA requirements

To properly work, DROPA requires 7 python libraries

- numpy (vers. 1.16.1) (Gold et al., 2015) is required to manage and process large matrices and arrays.
- tqdm (version 4.31.1) (available at https://github.com/tqdm/tqdm/tree/v4.31.1) was used to provide a progress bar of DROPA analysis.
- pandas (version. 0.24.1) (McKinney, 2010) is required to process large tables and to perform row or column operations
- intervaltree (version 3.0.2) (available at https://github.com/chaimleib/intervaltree) was used to compute genomic overlap between query peaks and gene reference.

- upsetplot (version 0.2.1)(Lex et al., 2014) was used to create upset plot results.
- matplotlib (version 3.0.3)(Hunter, 2007) was used to create plots like histograms and pie charts.
- argparse (version 1.4.0) was used to create the command-line interface.

Furthermore, to perform peak randomization, bedtools (see section 2.1.2.2) shuffle command is used.

2.2.2 DROPA performance evaluation

2.2.2.1 Influence of expression data metrics on DROPA

To assess DROPA performance using different gene expression unit, an experimental set of DRIP-seq peak (both available at GEO: GSE115957) and correspondent RNA-seq library were used. TPM and FPKM were computed from RNA-seq data using human RefSeq gene reference. DROPA analysis was made two times with default settings using hg19_Refseq as gene reference and either TPM or FPKM unit for expression data. Results of annotation obtained with different gene expression unit were compared counting how many peaks were annotated in the same way in both cases.

2.2.2.2 Assessment of DROPA performance

To assess DROPA performance using stranded data a DRIPc-seq peak dataset and correspondent RNA-seq library (both available at GEO: GSE70189) were used. DROPA analysis was launched with default settings, using hg19_UCSCgenes as genome reference, without considering strand of DRIPc-seq peaks. Strand of these peaks predicted by DROPA was then compared with the one of the original set.

2.2.2.3 DROPA comparison with existing tools

DROPA was compared with 3 broadly used peak annotation tool: PAVIS (Huang et al., 2013), UROPA (Kondili et al., 2017) and HOMER (Heinz et al., 2010). In all 3 comparison a set of DRIP-seq peak (available at GEO: GSE115957) and correspondent RNA-seq library were used. As each different tool has different degree of customization (limited gene reference selection, uneditable upstream/downstream dimension, etc.), DROPA settings were adapted to the one of the tools in comparison.

Comparing DROPA with HOMER, DROPA was launched with upstream/downstream region dimensions set to 1 kb and RefSeq gene reference. HOMER was launched with default settings. Comparing DROPA with PAVIS, DROPA was launched with default settings and UCSC_knowngene reference, while PAVIS was launched setting the Upstream/Downstream region to 5kb (same as DROPA default). Comparing DROPA with UROPA, DROPA was launched with default settings and Ensembl gene reference, UROPA was launched setting the Upstream/Downstream region to 5kb.

For each comparison, peak annotation results were analysed counting the number of intergenic peaks in each condition and the number of annotated to the same gene.

2.3 RNA-seq analysis

For each library, total cellular RNA was purified with the acid phenol method, quantified by UV absorbance and quality controlled by electrophoresis. RNA was then depleted of rRNA by Ribo-zero rRNA Removal Kit (Illumina) and libraries prepared with NEBNext Ultra Directional RNA library prep Kit for Illumina (NEB #E7420S) following manufacturer instructions.

All RNA-sequencing libraries produced by our laboratory were sequenced by Illumina HiSeq4000 platform (pair-end 2X150 bp) at Biodiversa S.r.l. (Rovereto,TN, Italy).

2.3.1 RNA-seq pipeline

All the RNA-seq libraries followed a standard RNA-seq workflow (Figure 2-2).

Quality checking, trimming and alignment sorting and indexing were performed in the same way as for DRIP-seq libraries (see section 2.1.1).

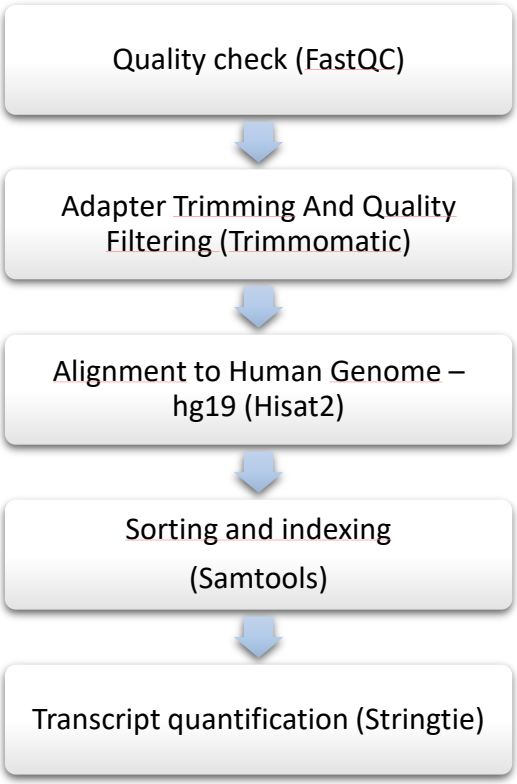

Figure 2-2 Workflow of RNA-seq analysis.

2.3.1.1 Hisat2

Hisat2 (version 2.1.0) alignment software for mapping next-generation sequencing reads (both DNA and RNA) of human genome (Kim et al., 2019). For each pair of read libraries were aligned on human genome (hg19) using the command line:

```
> hisat2 -p 8 --dta --fr -x indexes/hg19_UCSC/genome -1
trimmed.for.fastq -2 trimmed.rev.fastq -S output.sam
```

2.3.1.2 Stringtie

Stringtie (version 1.3.6) is a software that performs a RNA-seq alignments assembling into transcripts, or to quantify transcript using a gene reference (Pertea et al., 2015). For each alignment we used Ensembl GRCh37.87 gene reference and launched the following command:

```
> stringtie -B -e -G GRCh37.87.gtf -o library.gtf -A
library.abundance library.bam
```

where library.bam is the input alignment file and library.abundance is a table file containing gene quantification that will be used in differential expression analysis.

2.3.2 RNA-seq differential analysis and enrichment analysis

Differential expression analysis and gene set enrichment analysis were performed in R (version 3.5.3).

Library used in these analyses were:

- tximport (version 1.12.2) (Soneson et al., 2015): a R/Bioconductor package used to import reads count from Stringtie output
- DESeq2 (version 1.24.0) (Love et al., 2014): a R/Bioconductor package used to detect differentially expressed genes.
- tidyverse (version 1.2.1): a suite of R libraries that allow to manipulate large tables and matrices, compute row and column operations and create plots.
- fgsea (version 1.10.1) (Sergushichev, 2016): a R/Bioconductor package that implements an algorithm for fast gene set enrichment analysis, similar to GSEA (Subramanian et al., 2005).
- pathwiew (version 1.24.0) (Luo and Brouwer, 2013): A R/Bioconductor package used to map and render a RNA-seq data on relevant pathway graphs.

Stringtie transcript estimations were imported in R and converted to read count using tximport library. Differential expression analysis, together with PCA analysis was performed using DESeq2 with default settings and for each contrast. Differentially expressed genes were selected by adjusted p-value <0.05.

Gene Set Enrichment analysis was performed using fgsea library and MSigDB gene set database (version 6.2) with default settings. As input, summary result table of DE genes obtained with DESeq2 were used. Comparison of GSEA results between different treatment condition and between different cell lines was performed using tydiverse library.

Pathway visualization of RNA-seq data was performed using Pathview library with default settings.

2.4 PanCancer analysis

To analyse TCGA PanCancer data, in addition to libraries listed in 2.3.2, were used:

- survminer (version 0.4.6) was used to create survival plots.
- GSVA (version 1.32.0) (Hänzelmann et al., 2013) was used to perform single sample GSEA analysis.

Pre-computed data about copy number variations (CNVs), mutations and gene expression were processed and maintained by the Ciccarelli group at The School of Cancer Studies of King's College London and The Francis Crick Institute, as part of the Network of Cancer Genes Database (Repana et al., 2019).

To compute ratio of mutation for each query gene we used CNVs an mutation data and categorized mutations in five groups as reported in section 3.2.3. Computation of mutation rates and histogram plotting were performed using tidyverse and R base scripts.

To compute gene expression alteration between cancer and matched normal tissues, gene expression data were used. For each cancer type, only samples which had expression data for both cancer and normal tissue were selected. Computation and boxplots plotting were performed using tidyverse and R base scripts.

To compute survival analysis, we selected, for each cancer type, the 33% of samples with the higher and lower expression of the query gene. Survival plot and p-values were produced using survminer library.

To compute correlation between gene expression and enrichment scores, first we computed gene set enrichment score for each tumour sample using ssGSEA function from GSVA library. Then, for each query gene and for each cancer type, we computed spearman correlation and produced heatmaps using tidyverse and R base custom scripts.

3. RESULTS

This chapter will be divided in two main sections.

The first one will report on a study in which I have investigated R-loop dynamics in U2OS cancer cell lines treated with G4-binders and developed a new tool for annotation of R-loop peak allowing a finest gene annotation.

The second part will report on a study in which using RNA-seq data, I have observed an innate immune response induction in MCF-7 cells caused by treatment with a specific G4 binder, PDS. Moreover, the result of an analysis of cGAS/STING pathway genes expression in The Cancer Genome Atlas (TCGA) dataset and their correlation with immune response induction in tumour tissue will be reported as well.

The main results of the first part (section 3.1) are included in two publications (De Magis et al., 2019, Annex 1; Russo et al., 2019, Annex 2) while the results of the second part (section 3.2) are currently unpublished.

3.1 Part I: R-loop dynamics in U2OS cells treated with G4-binders

3.1.1 A pilot experiment

To test whether G4-binder treatments can affect R-loop formation in U2OS cancer cells we performed a pilot DRIP-seq sequencing experiment. The pilot data have not been published as they were intended only to get an evidence of G4 binder effects on R-loop levels and to define the analytic pipeline of DRIP-seq reads. The experimental design consisted of three DRIP-seq libraries representing three different experimental conditions:

- U2OS cells treated with BRACO G4-binder for 5 minutes;
- U2OS cells treated with FG G4-binder for 24 hours;
- U2OS cells untreated as control condition;

All three pair-end libraries were quality filtered and adapter trimmed using Trimmomatic tool and were aligned on the human genome (hg19 version) using BWA software. Next, by using Samtools suite, only uniquely mapped and properly paired reads were considered, and all the libraries were downsampled to the smallest one to normalize genomic signals.

The results of the alignment (Table 3-1) show that rates of properly-mapped and paired reads range between 64% and 76%, due to the high number of duplicated reads.

Library	**Library dimension (n. of reads)**	**Mapped reads (% of total)**	**Properly paired reads (% of total)**
BRACO-treated	47,441,573	66.93%	66.24%
FG-treated	60,605,572	76.92%	76.36%
Control	64,949,827	65.73%	64.56%

Table 3-1 Results of library alignment on human genome(hg19) in the pilot experiment.

Then, peak calling analysis was performed for each library using peak caller using default settings and a mock library of genomic DNA randomly fragmented as input control. This tool also performs a differential analysis of signal levels between treatment and control condition.

Results of peak calling (Table 3-2) show that, in both treatment conditions, a high number of peaks with a signal level ratio over 1.5X (gain peaks) under 0.66X (loss peaks) were identified, with a prevalent presence of loss regions.

Condition	N. DRIP-seq peaks	Gain Peaks	Loss peaks
BRACO-treated vs. Control	86,506	12,546	40,267
FG-treated vs. Control	106,378	16,543	31,842

Table 3-2 Results of peak calling analysis using ODIN in the pilot experiment.

Annotation of these peaks on the genome using PAVIS (Huang et al., 2013) shows that as expected most of these peaks are intragenic, and that loss peaks are particularly enriched in transcription start site region (Figure 3-1).

-

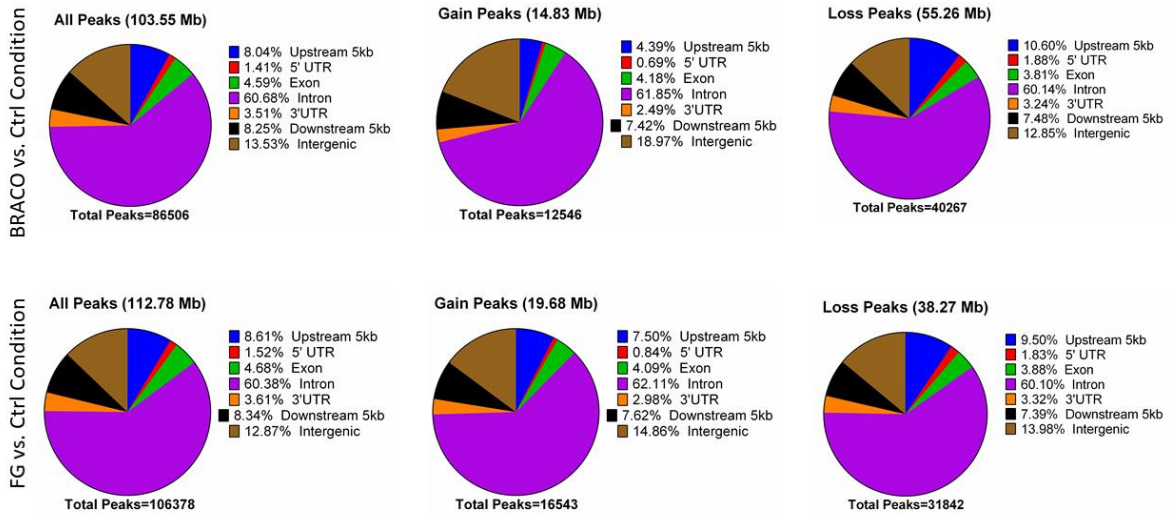

Figure 3-1 Distribution of Total, gain and loss peaks over genic features using PAVIS in the pilot experiment.

To investigate R-loop interaction with G4 structures, we compared our R-loop peaks to G4 structures determined by other labs. In particular, I used a genomic G4 structures dataset of 716,310 G4 loci derived with a polymerase stop assay and NGS using genomic DNA (Chambers et al., 2015). Although *in vivo* G4 datasets were available (Hänsel-Hertsch et al.,

2016; Kouzine et al., 2017), we discarded these from the analysis since none of these dataset were related to our cell line. In fact, *in vivo* G4 datasets show a consistently lower number of structures compared to *in vitro* dataset. Since G4 loci mapped with G4-ChIP in two different cell lines showed a certain degree of tissue specificity, we opted for a less specific dataset which includes all possible G4 structures on the genome.

I performed a colocalization analysis of G4 structures and R-loop peaks using ColoWeb tool. Results of this analysis (Figure 3-2) showed that there is a higher level of colocalization between R-loop regions and G4 loci than between G4 structures and R-loop peaks randomized all over the genome.

While these preliminary findings were from a single pilot experiment, they were promising and suggested to perform a complete experiment with biological replicates. Furthermore, to fully investigate R-loop/G4 interaction, we need to consider R-loop strandness. Since DRIP-seq methodology does not provide this type of information, section 3.1.2 will illustrate how to predict strandness of R-loop peaks using gene expression information.

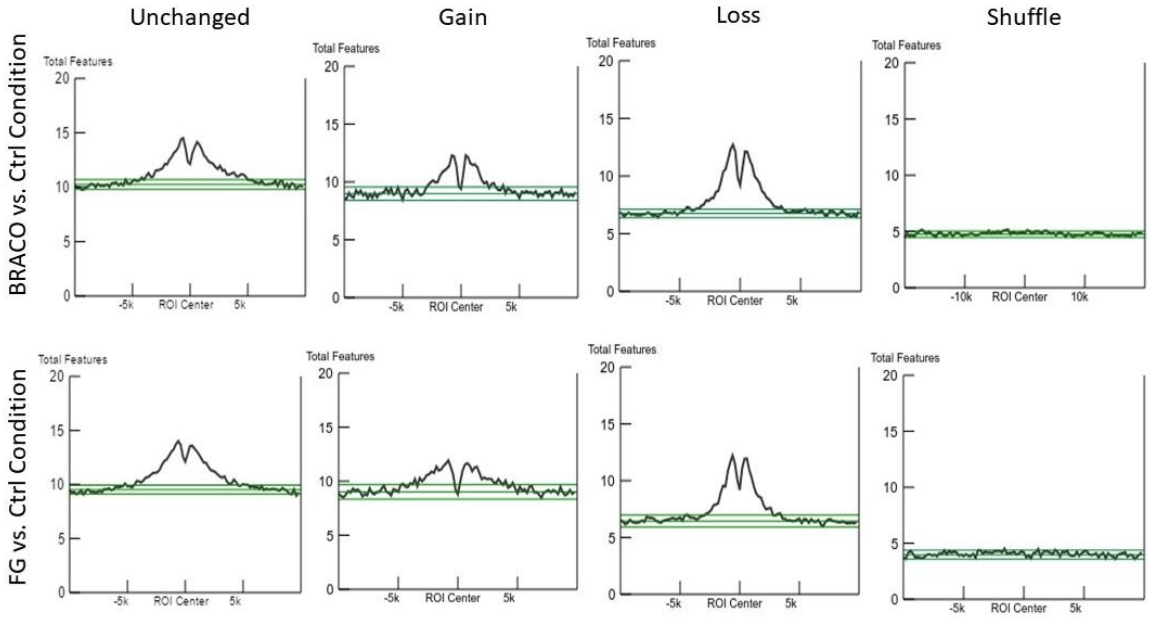

Figure 3-2 Distribution of G4 structures over R-loop peaks. For each R-loop peak, a window of 20kb is calculated from centre position. Shuffle category contains R-loop peaks randomized all over the genome. Green lines indicate estimation of normal variance above and below the background level.

3.1.2 DROPA: DRIP Optimized Peak Annotator

DRIP-seq peaks often have a dimension in the order of kilobases. This means that these peaks in many cases may overlap more than one gene region or even different genes. Moreover, DRIP-seq peaks have no strand information. However, that is essential to assess the overlapping of R-loops with G4-structures as R-loop is constituted of a hybrid duplex involving one DNA strand leaving the other one in a single-stranded status. As a G4 structure in the latter strand only would be compatible with an R-loop in the same genomic region, therefore it is important to know which DNA strand is annealed to the RNA in the R-loop peak. Then, as R-loops are mainly co-transcriptional events, we developed a new tool for peak annotation optimized for DRIP-seq peaks, which uses gene expression information to perform a finest peak annotation considering the DNA strand of hybrid formation.

DROPA (DRIP Optimized Peak Annotator) is a command line tool written in Python language. It requires three files as input:

I. The query peak file with the result of peak calling in BED format;
II. A gene reference folder that contains all the data about gene features (5'UTR, 3'UTR, exon, intron) in BED format. In the final version of this tool we added reference set for *Homo sapiens* (hg19) and for *Mus musculus* (mm9 and mm10), but they can be easily created for every genome of interest by the user;
III. A 2-column table containing gene expression data with the name of each gene and its normalized expression value (FPKM, TPM, etc.);

DROPA code is composed by six main scripts, as shown in Figure 3-3:

1) *PeakOverlap*: the script performs an overlap analysis using query peak file and gene reference. As output it provides two BED files, one containing intergenic peaks that are excluded from further analysis and one reporting all the overlaps between a query peak and reference genes;
2) *CheckExpression*: it represents the major change comparing to other peak annotation tools. It takes in account gene expression information to uniquely annotate each peak to a gene. When a query peak is overlapping only one gene it is easily assigned to it, while in case of 2 or more overlapping genes the most expressed one is chosen. In addition, to refine the annotation, an expression threshold can be set. In the extreme

case in which all the genes are not expressed or their expression is below threshold, query peak is assigned to the gene that has the higher overlap;

3) *FeatureAssign*: it assigns a query peak to all the overlapped gene features (upstream/downstream region, intron, exon, UTR (Untranslated Region) regions);
4) *TableCreator*: this script writes a summary table with an entry for each peak and all the information about annotated genes and gene features;
5) *RandPeak*: it uses "shuffle" Bedtools command to perform query peaks randomization all over the genome and then relaunches 1) to 4) scripts. The output is used to perform enrichment score over randomized peaks;
6) *SummaryPlot*: it uses all produced data to plot annotation results.

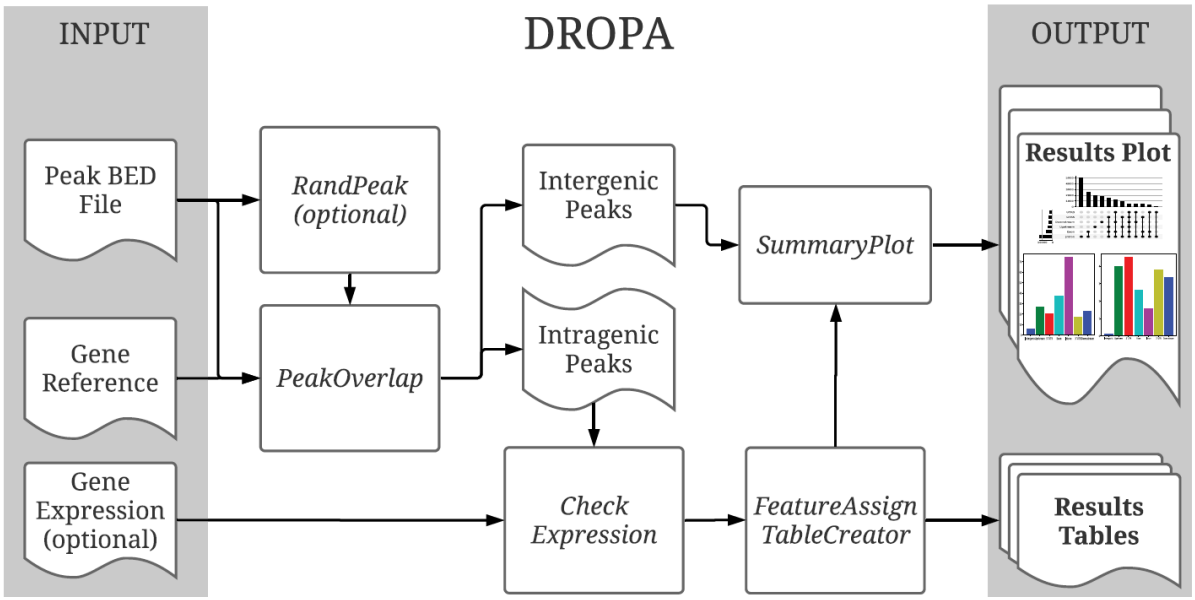

Figure 3-3 Schematic representation of DROPA workflow. Adapted from Russo et al., 2019.

To prove the efficiency of the new tool, I performed different tests. First, I assessed whether using different expression values (TPM or FPKKM) leads to different annotation results. The annotation analysis was performed with default settings using the same query peak dataset and relative RNA-seq data, which was used to compute expression level as TPM and FPKM. Results of this test were good as they demonstrated that 98.7% of query peaks were annotate to the same gene using either TPM or FPKM. Then, we assessed the efficiency of DROPA in annotating peaks using a stranded DRIPc-seq query peak dataset and his relative

RNA-seq data (Russo et al., 2019). In this analysis we compared how many DRIPc-seq peaks are annotated to the correct strand using DROPA. Results of this test show that most of the peaks (88.6%) are assigned correctly. Furthermore, of the remaining 11.4% of wrongly annotated peaks, 5.05% of them are localized in the opposite strand of the same locus of another DRIPc peak, while 3.53% are localized outside of the gene region but within 5000 bases from transcription start site (TSS) end termination site (TTS). Since antisense transcription is known to occur at these sites, we hypothesized that most of wrongly assigned peaks are actually antisense R-loop, that cannot be recognized by DROPA. However, when we consider only transcribed regions of a gene, DROPA efficiency reaches 93.8%.

Finally, we tested DROPA efficiency compared to 3 common annotation tools (HOMER, PAVIS and UROPA). Since each tool had a different degree of customization, DROPA was launched adapting his settings to the one each tool in analysis. Results of the analysis shows that DROPA annotates fewer peaks as intergenic than all the three tested tools. Furthermore, a high percentage of peaks are annotated to different genes by DROPA compared to each tested tool. As R-loop formation is mainly co-transcriptional, and DROPA is the only tool that perform annotation based on gene expression, we argue that DROPA performs a more robust annotation, with less false positives compared to other general standard tools. Further information on DROPA behaviour and his benchmark comparison have been published (Russo et al., 2019). Benchmark comparison showed that DROPA perform analysis in times that are comparable with those of other tools and without high system requirement.

Since this tool was developed to perform a better annotation of DRIP-seq peaks, and its good performance was extensively demonstrated, I used it in subsequent analysis.

3.1.3 DRIP-sequencing results of U2OS cells treated with G4-binders

As the pilot data (section 3.1.1) lacked biological replicates but results about R-loop/G-quadruplex interaction were promising, we repeated R-loop mapping via DRIP-seq with a new experimental design. Since we had experimental evidences by immunofluorescence microscopy of an increase of R-loop structures in U2OS cells (De Magis et al., 2019) using G4-binders including FG and pyridostatin (PDS), DRIP-seq libraries were prepared from the following samples:

i. U2OS cells treated with FG (2 replicates) for 5 minutes.
ii. U2OS cells treated with PDS (2 replicates) for 5 minutes.
iii. U2OS cells untreated as control condition (2 replicates).

For each condition, a negative control of DRIP library was prepared by treating genomic DNA with *E.coli* RNAseH, which specifically degrades RNAs annealed to DNA strands only, before immunoprecipitation with Ab S9.6.

Following the analytic pipeline described in 3.1.1, reads were trimmed and quality filtered with Trimmomatic, aligned on hg19 reference genome with BWA, filtered for duplicates and downsampled with Samtools.

The results of the alignment show that properly mapped and paired reads rate ranges between 63% and 73%, except for Control Rep.1 (43.82%) and FG-treated+RNAseH (49.06%) (Table 3-3).

Library	Library dimension (n. of reads)	Mapped reads (% of total)	Properly paired and filtered reads (% of total)
Control Rep.1	113,248,498	76.41	43.82
Control Rep.2	92,777,074	76.05	66.16
Control + RNAseH	143,465,986	89.73	66.89
FG-treated Rep.1	96,765,498	77.69	66.65
FG-treated Rep.2	82,192,684	79.64	71.04
FG-treated + RNAseH	103,419,848	87.59	49.06
PDS-treated Rep.1	100,312,608	77.23	63.68
PDS-treated Rep.2	83,977,062	81.98	73.84
PDS-treated + RNAseH	91,652,406	88.75	70.86

Table 3-3 Results of library alignment oh human genome(hg19).

Using Deeptools suite, correlation between libraries was computed to assess the consistency of replicates. As shown in Figure 3-4, DRIP-seq libraries show a high correlation coefficient between replicates, as RNAseH treated libraries between them. DRIP-seq libraries show a high correlation coefficient also between different condition of treatment and control, suggesting that R-loop signal is generally localized in specific loci without no major changes

between the treatment condition. On the other hand, RNAseH treated samples show a sensibly lower level of correlation with DRIP-seq libraries, suggesting that R-loop signal was successfully suppressed.

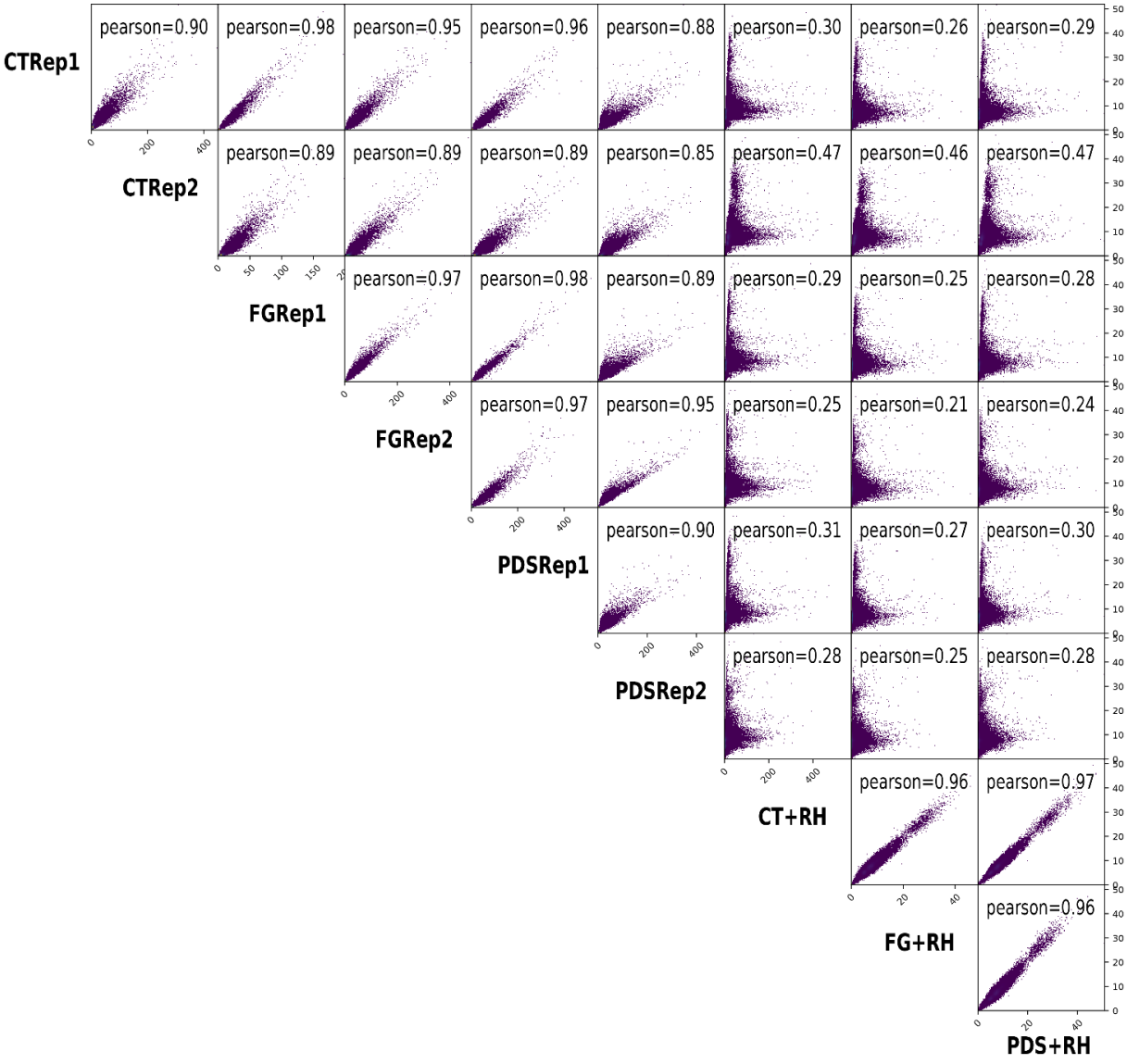

Figure 3-4 Scatter plot showing correlation between DRIP-seq libraries read counts. Each dot represents a 10 kilobases bin of human genome (hg19). Inside each box Pearson correlation coefficient is reported.

3.1.4 DRIP-seq peaks are strongly induced in U2OS cells treated with G4-binders

Peak calling was performed with MACS2 using default settings for each library, and then we selected for further analysis only peaks present in both replicates but absent in RNAseH-

treated samples. To assess the consistency of data between replicates, we calculated correlation coefficients of such DRIP-seq peaks signal using Deeptools (Figure 3-5). The results showed a high correlation between each of two biological replicates but not with RNAseH treated negative control, supporting a high consistency of observed genomic R-loop peaks (Figure 3-5).

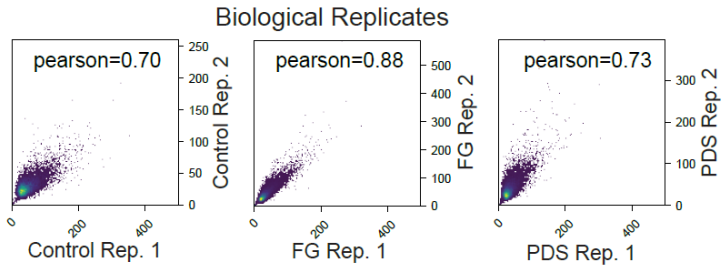

Figure 3-5 Scatter plot showing correlation between DRIP-seq libraries read counts. Each dot represents a DRIP-seq peak. Inside each box Pearson correlation coefficient is reported. Adapted from De Magis et al., 2019.

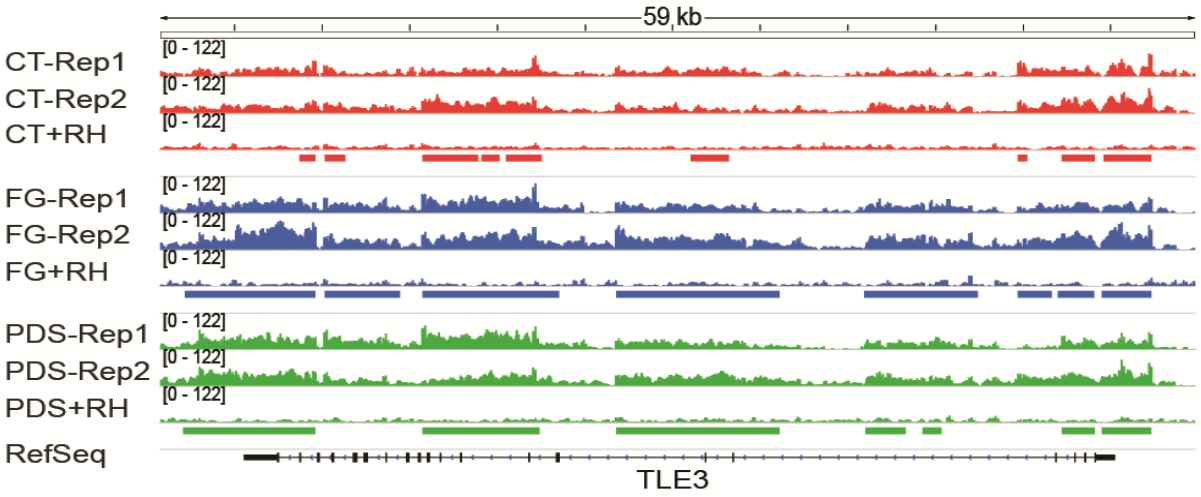

Figure 3-6 Genomic signal of DRIP-seq for control (red) FG-treatment (blue) and PDS (green) condition at TLE3 locus. Coloured bars indicate DRIP-seq peaks called by MACS2. Adapted from De Magis et al., 2019.

Results of peak calling (Table 3-4) show that in both FG an PDS-treated condition there is a strong increase of R-loop peaks in terms of number of peaks and genome covered.

Condition	N. Consensus DRIP-seq peaks	Covered genome (Mb)	Covered genome (%)
Control	18,262	77.564	2.5
FG-treated	37,068	150.802	4.8
PDS-treated	37,974	158.646	5.1

Table 3-4 Results of peak calling analysis using MACS2.

Then, I annotated R-loop peaks in sample using DROPA tool (section 3.1.2). As expected, R-loop peaks are mostly localized at genic level, with a consistent enrichment in 5' and 3' regions of genes compared randomized peaks.(Figure 3-7).

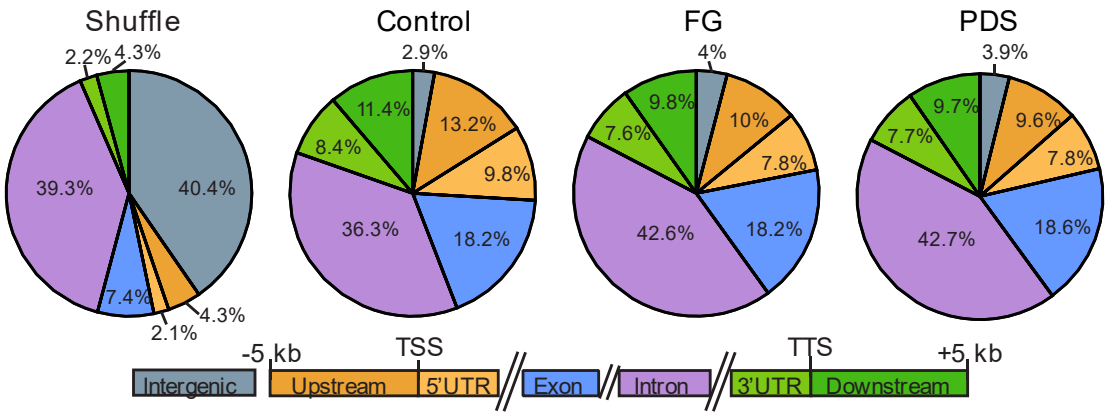

Figure 3-7 Distribution of Control, FG and PDS DRIP peaks across the genic features. On the left, "shuffle" shows the distribution of randomly shuffled peaks over the entire genome. Adapted from De Magis et al., 2019.

Since R-loop are also associated with transcription initiation event, we measured R-loop signal level at promoter regions. Promoter regions were divided in four groups on the base of gene expression level (calculated using RNA-seq data). Results of this analysis (Figure 3-8) showed that R-loop changes are positively correlated with gene expression level and the presence of CpG island, which are prevalent in highly expressed genes.

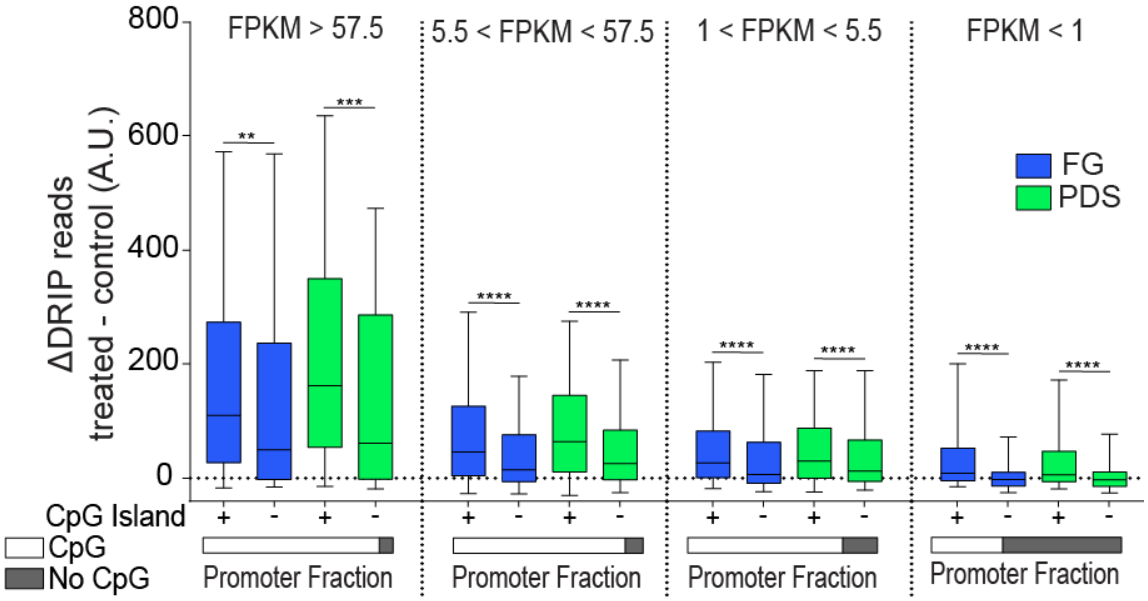

*Figure 3-8 DRIP-seq read counts at active TSS level. Analysed regions are +-2,000 bp upstream and downstream from TSS. Genes are divided into four categories based on gene expression as established by RNA-seq data and indicated at the top of the graphs. As indicated below, TSS promoters with CG islands constitute 92.9, 90.6, 81.8, and 37.4% for the four gene sets from left to right, respectively. Statistical test: Kolmogorov–Smirnov test. **P < 0.01, ***P < 0.001, ****P < 0.0001. Adapted from* De Magis et al., 2019.

We observed spreading of R-loop peaks both in terms of elongation of pre-existent R-loops and formation of new ones. To detect local changes of R-loop peaks intensity, a differential analysis of using edgeR was performed. Results (Table 3-5) showed that in both FG and PDS treatment there is a large number of regions in which there is an increase in R-loop signal (Gain peaks), and only few in which there is a lower R-loop signal (Loss peaks). Gain peak localization was particularly enriched at the 3' regions of genes (Figure 3-9). However, we did not observe any specific correlation with G4 structures presence.

Condition	N. Gain DRIP-seq peaks	N. Loss DRIP-seq peaks
FG-treated vs. Control	4,411	149
PDS-treated vs. Control	9,881	272

Table 3-5 Results of differential binding analysis of DRIP-seq libraries.

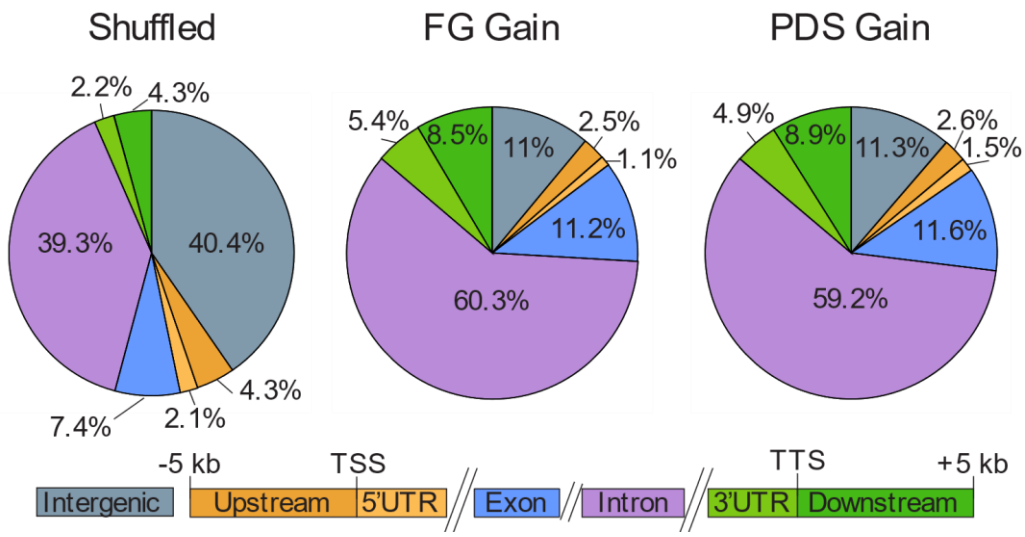

Figure 3-9 Distribution of FG gain and PDS gain DRIP peaks across the genic features. On the left, "shuffle" shows the distribution of randomly shuffled peaks over the entire genome. Adapted from De Magis et al., 2019.

Thus, we assessed whether the increase of the length of pre-existing R-loops was a significant event. We identified a high number of peaks that were present in both control and treated samples (13,539 and 13,316 common peaks for FG and PDS, respectively) mapping at the same genomic locus. Among these common peaks 1000 and 639 peaks, for FG and PDS, respectively, showed a statistical-significant increase of the length in the treated samples (Figure 3-10).

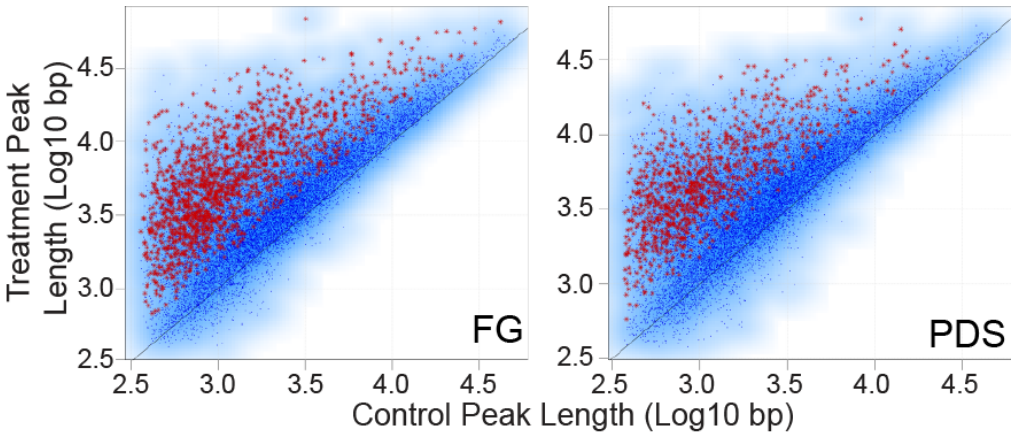

Figure 3-10 Scatter plots of DRIP-seq peaks lengths of common peaks between control and treatment condition. Red asterisks highlight extended peaks with a length fold change >1.5 and P-val< 0.05. Tests used: moderated t test from the limma R package. Adapted from De Magis et al., 2019.

Interestingly, an annotation analysis of the G4-binder-extended peaks showed that they were particularly enriched at 3'-ends and in the body of genes (Figure 3-11).

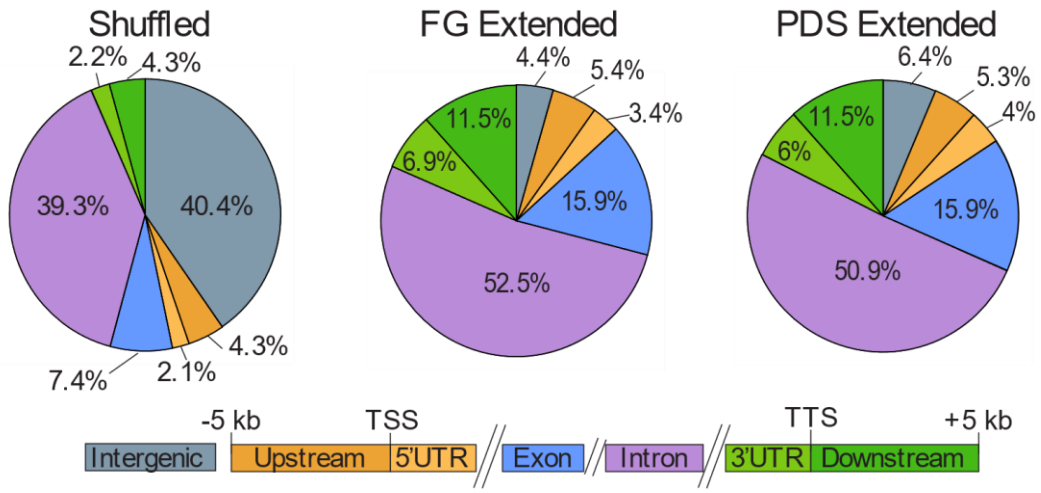

Figure 3-11 Distribution of FG and PDS extended DRIP peaks across the genic features. On the left, "shuffle" shows the distribution of randomly shuffled peaks over the entire genome. Adapted from De Magis et al., 2019.

3.1.5 R-loop interplay with G4 structures in U2OS cells

To investigate R-loop interactions with G4 structures, we used the same G4 structure dataset as in section 3.1.1. We observed a good overall correlation between the localisations of R-loop peaks and G4 structures in both control and treatment conditions (Figure 3-12), consistently with a potential structural co-existence of both structures at same loci.

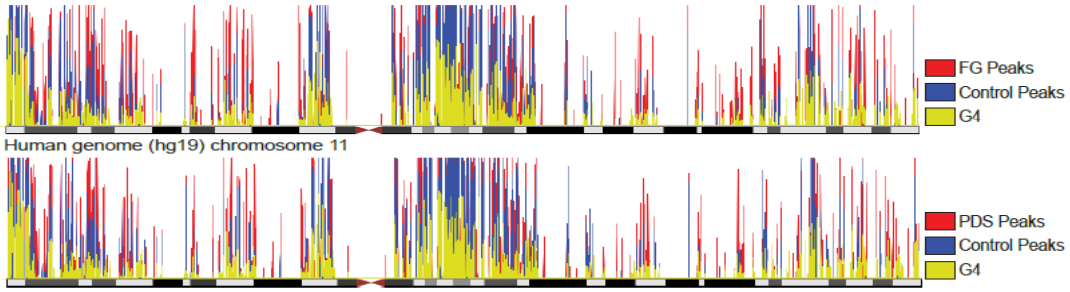

Figure 3-12 Representative distribution of DRIP-seq peaks and G4 loci across human chromosome 11. Only transcribed gene regions are considered. Pearson correlation coefficient between FG and PDS

peaks and G4 loci all over the genome is 0.60 and 0.53 respectively. Pearson correlation coefficient between FG and PDS peaks and Control peaks all over the genome is 0.81 and 0.78 respectively. Adapted from De Magis et al., 2019.

Since DRIP-seq peaks have no strand information, we used the template strand of the gene annotated by DROPA as the strand forming the DNA:RNA hybrid duplex. This allowed to distinguish G4s on the base of their position relative to the hybrid of R-loops, splitting them in two groups: G4s that are localized in the displaced single DNA strand of the R-loop and G4s that are localized in the DNA strand forming the hybrid of the R-loop. Thus, the former G4s would be compatible with R loops, whereas the latter G4s were not structurally compatible with hybrid (R loop) formation.

As we noticed that in many loci R-loop extensions co-occurred with G4 structures (Figure 3-13), we computed the enrichment of this co-occurrence versus the expected event calculated with the same number of observed R-loop peaks, randomized over genic regions and normalized for length and genic localization.

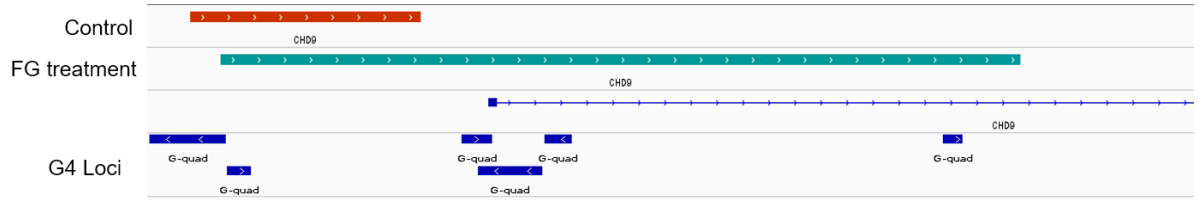

Figure 3-13 Representative locus showing extended R-loop peak in control and FG treatment condition with G4 motif presence in the extended region of the peak. Snapshot taken with Integrative Genomics Viewer.

Then, I compared the enrichment score of observed R-loop peaks with a "matched" peak set, which is composed of unchanged common R-loop peaks of similar sizes, randomly selected. The results show a significative enrichment of G4s at the displaced strand of the R-loop for both FG and PDS, whereas no enrichment was detected for G4s s at the template DNA strand of the R-loop (Figure 3-14). In many cases G4s were present in both DNA strands of the R-loop, and a significant enrichment compared to matched was observed only for FG treated condition.

I also noticed that the absence of G4 structures within R-loop extensions showed no differences between extended and matched peak datasets for PDS condition, while in case of FG there is a significative depletion. The lack of significance of G4/R-loop association in certain cases (see PDS vs FG above) can be due to the tendency of physiological R-loops to form in regions with an high GC skew and G4-promoting sequences already, which therefore lead to a high overlapping rate of R-loops with G4 structures at basal levels.

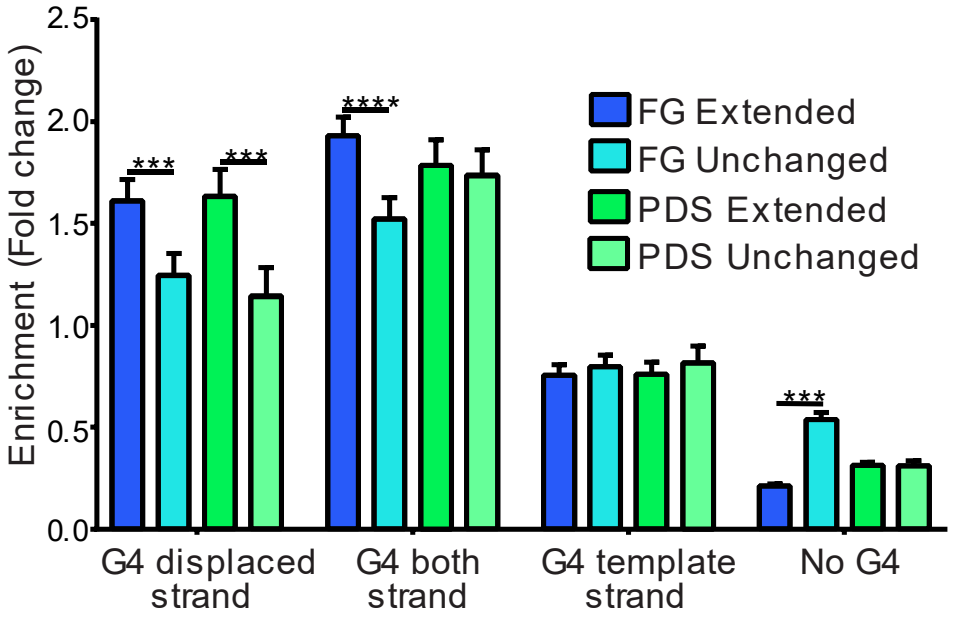

*Figure 3-14 Enrichments over expected of G4 motifs in extended regions of extended peaks a set of unchanged peak, matched to extended ones. Only extended peaks in genic regions were considered for the analysis. Test used: Kolmogorov–Smirnov test. ***P < 0.001, ****P < 0.0001. Adapted from* De Magis et al., 2019.

The above findings, together with immunofluorescence experiments (De Magis et al., 2019), demonstrate that G4-binder treatments induce an increase of R-loop structures in U2OS cells. Furthermore, bioinformatic analysis of DRIP-seq suggests that this increase can be at least partially explained by G4 structure presence at the displaced strand of the R-loop, likely serving as a stabilizing factor of the observed R-loop extensions.

3.2 Part II: Innate immune response induction in cancer cells and cGAS/STING pathway genes in cancer tissues

3.2.1 Gene expression analysis of MCF7 cells treated with pyridostatin

The bioinformatic analysis of R-loop/G4 interplay described in this Thesis has been part of a study that included also wet experimental findings(De Magis et al., 2019). In the study, we also found that one of the effects of pyridostatin in U2OS cells was the induction of micronuclei, which can be a source of cytosolic DNA triggering an innate immune response via the cGAS/STING pathway (see also Introduction). Since in U2OS cells, STING appears to be inactive (Deschamps and Kalamvoki, 2017), we then asked the question of whether PDS can induce an innate immune response via STING pathway in the breast cancer MCF-7 cell line, which expresses an active STING and produces micronuclei upon PDS treatments (unpublished).

In order to establish if PDS can induce an immune response in cancer cells, we performed RNA-seq experiment in MCF-7 cells upon PDS treatment. Four biological replicates were prepared for each of the following samples and experimental conditions:

- MCF-7 cells untreated (Control_t0)
- MCF-7 cells treated with PDS + 4 days of recovery (PDS_t4) in drug free medium
- MCF-7 cells untreated + 4 days in drug free medium (Control_t4)

We obtained 12 RNA-seq libraries, that were processed for read trimming with Trimmomatic, and aligned on the human genome (hg 19 version) using Hisat2 software. Results of alignments (Table 3-6) showed a high percentage of properly paired reads in all libraries.

Library	Library dimension (n. of reads)	Mapped reads (% of total)	Properly paired reads (% of total)
Control_t0 Rep.1	102,619,128	98.72	95.07
Control_t0 Rep.2	112,140,222	98.67	95.04
Control_t0 Rep.3	102,051,560	98.73	94.80
Control_t0 Rep.4	77,983,079	98.95	95.25
Control_t4 Rep.1	102,056,415	98.68	94.93
Control_t4 Rep.2	157,098,969	98.79	95.06
Control_t4 Rep.3	70,821,281	98.77	94.90
Control_t4 Rep.4	78,405,913	98.94	95.41
PDS_t4 Rep.1	99,923,734	98.57	94.74
PDS _t4 Rep.2	102,595,656	98.60	94.91
PDS _t4 Rep.3	77,190,875	98.79	94.50
PDS _t4 Rep.4	79,943,162	98.75	95.35

Table 3-6 Results of RNA-seq libraries alignment oh human genome(hg19).

Gene expression was quantified using Stringtie software, followed by a differential expression analysis performed using DEseq2 R package. Results of differential analysis showed many overexpressed genes (Table 3-7) in all the contrasts we have tested (PDS_t4 vs. Control_t4, PDS_t4 vs. Control_t0 and Control_t4 vs. Control_t0). Notably, while PDS_t4 vs. Control_t0 contrast has the highest number of differentially expressed genes, we also observed significant changes in expressed genes in the Control_t4 vs. Control_t0,, suggesting that some of the differences in gene expression that we observed in PDS-treated cells may not be due only to treatment with PDS.

Condition	N. overexpressed genes	N. underexpressed genes
PDS_t4 vs. Control_t4	189	81
PDS_t4 vs. Control_t0	426	356
Control_t4 vs. Control_t0	135	128

Table 3-7 Results of DEseq2 differential expression analysis.

As shown in Figure 3-15, top 100 most significantly up- and downregulated genes show a good consistence between replicates. Notably, *IFNB1* (Interferon beta gene) and other interferon induced genes (e.g. *IFIT1* and *IFIT3*) show a specific expression in PDS treated samples only.

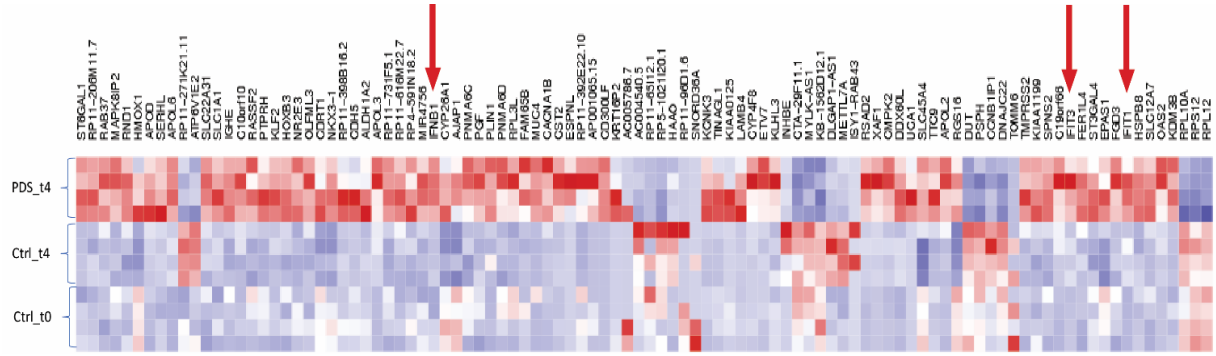

Figure 3-15 Top 100 differentially expressed gens in PDS_t4 vs. Ctrl_t4 contrast and expression values for each library. Expression values are reported as Z-score. Blue = -3, Red =3. Red arrows highlight genes cited in the main text.

Then, we performed gene set enrichment analysis (GSEA) to investigate if overexpressed genes where related to particular pathways. Results of GSEA showed that in PDS_t4 vs. Control_t4 condition (Figure 3-16), most upregulated MsigDB hallmark gene sets (Subramanian et al., 2005) are related to interferon and inflammatory response, suggesting that PDS treatment has an effective role in immune response induction.

On the other hand, most downregulated pathways are related to cell proliferation (*Myc* and *E2F* targets, G2/M checkpoint), underlying the DNA cleavage induction by PDS treatment.

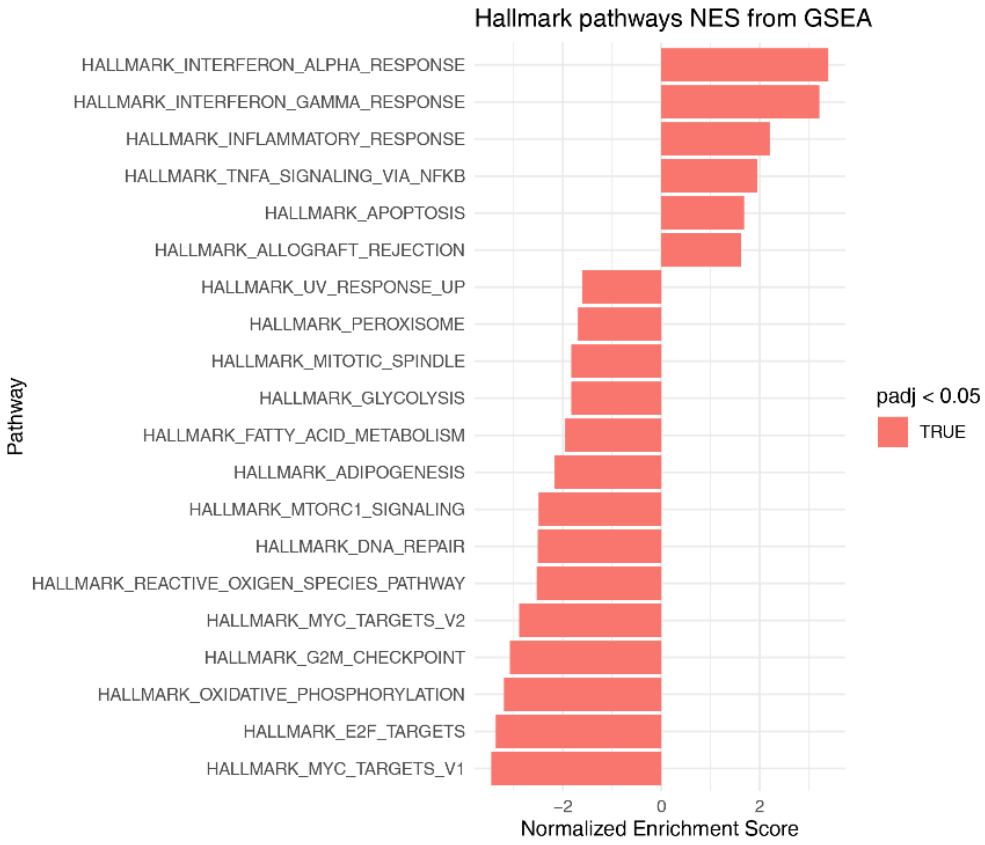

Figure 3-16 Bar plot with GSEA analysis results of PDS_t4 vs. Ctrl_t4 using MsigDB hallmark gene sets. Only top 20 gene sets ranked by adjusted p-value are reported.

Using the control at time 0 (Control t0 sample) as reference, we performed a comparison between the time-dependent effects of PDS-treated and untreated cells.

Interestingly, the data showed that the most up-regulated gene sets are the interferon alpha and gamma response in both treated and untreated cells, however the normalized enrichment scores are higher in the PDS_t4 vs. Control_t0 contrast (3.38 and 3.12, respectively) than in Control_t4 vs. Control_t0 contrast (2.73 and 2.38, respectively) (Figure 3-17). Most of downregulated pathway show the same result, with a lower normalized but still significant enrichment score in Control_t4 vs. Control_t0 contrast.

Since Hallmark interferon alfa response (composed by 97 genes) share 73 genes with Hallmark interferon gamma response (composed by 200 genes), it is likely that the similar result between these two signatures is due to the high redundancy of genes.

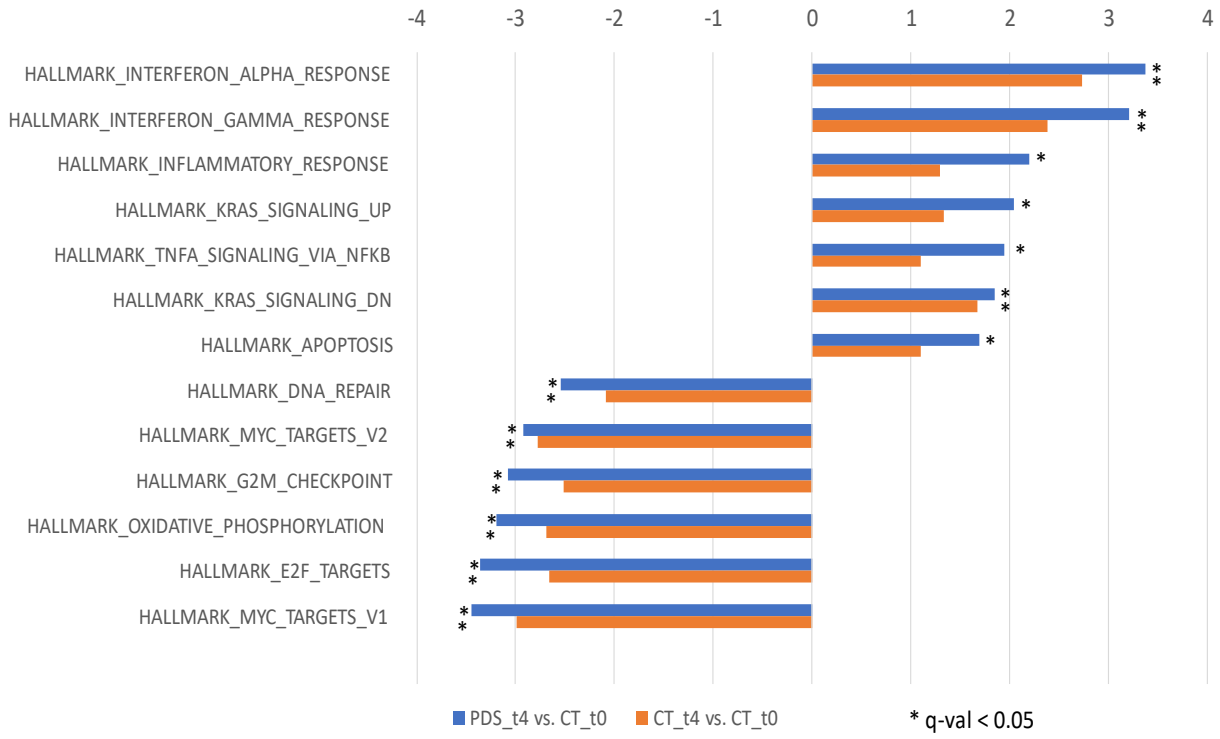

Figure 3-17 Bar plot with GSEA analysis results of PDS_t4 vs. Ctrl_t0 (blue) and Ctrl_t4 vs. Ctrl_t0 (orange) contrasts using MsigDB hallmark gene sets. Only Top13 gene sets ranked by adjusted p-value are reported.

We further investigated the comparison between PDS-treated and untreated cells using MsigDB Reactome gene sets. The data showed a result in agreement with Hallmark dataset: most up-regulated pathways are related to interferon alpha/beta and interferon signalling and are significant for both conditions (Figure 3-18). Regarding down-regulated pathways, as in Hallmark dataset regards DNA replication and cell cycle progression (Figure 3-18).

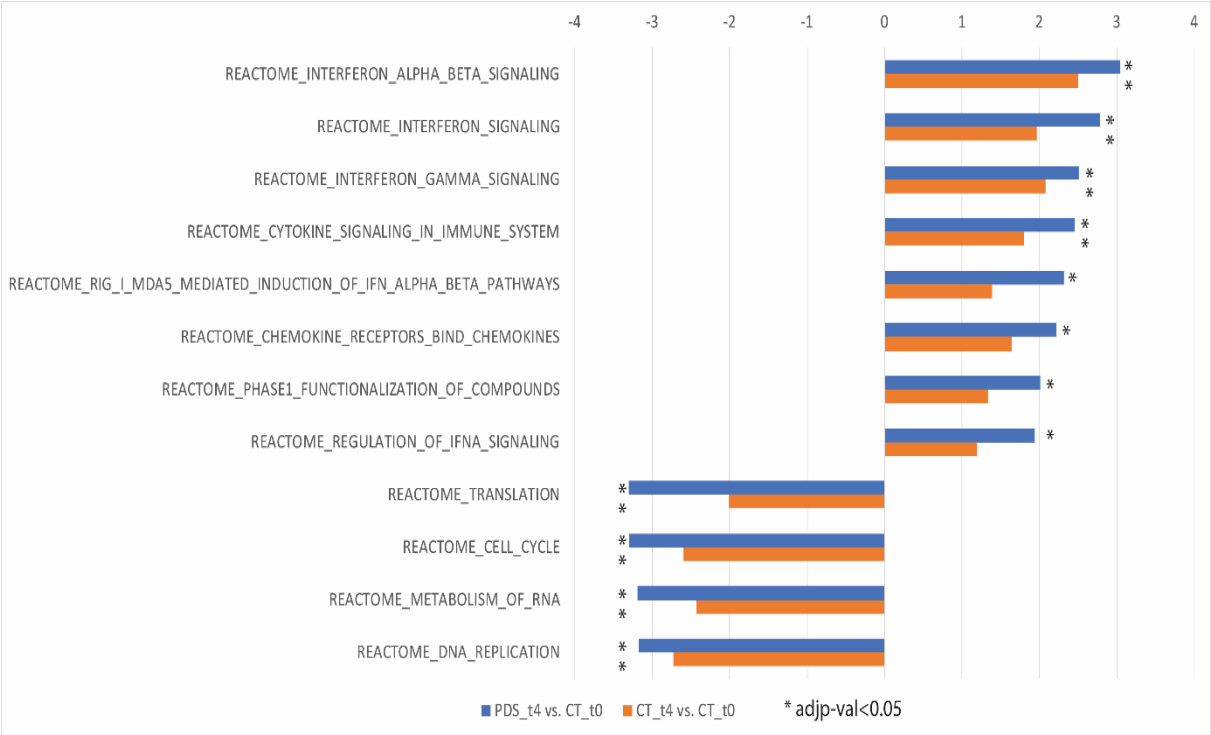

Figure 3-18 Bar plot with GSEA analysis results of PDS_t4 vs. Ctrl_t0 (blue) and Ctrl_t4 vs. Ctrl_t0 (orange) contrasts using MsigDB Realtime gene sets. Only Top12gene sets ranked by adjusted p-value are reported.

Thus, we directly tested whether genes that belong to the "interferon alpha response" set showed a different expression level between PDS_t4 and Control_t4. Comparison of Wald statistic between PDS_t4 vs. Control_t0 and Control_t4 vs. Control_t0 contrasts showed that while many genes of the "interferon alpha response" gene set are indeed overexpressed in Control_t4 vs. Control_t0 contrast, the overexpression is significantly higher (p-val<10E-16, Kolmogorov Smirnov test) in the PDS_t4 vs. Control_t0 contrast (Figure 3-19). We observe the same effect whether we take in account other gene sets, such as "Reactome_Alpha_Beta_Signaling" (Figure 3-19).

These findings are likely explained by the fact that we observe a small but consistent increase of micronuclei number after 4 days of cell growth in drug-free medium, even though the micronuclei number is indeed significantly higher in PDS treated cells at the same time of cell growth (unpublished data). This may probably explain the increase of interferon response genes in control, untreated cells during cell culture, which PDS treatment can significantly boost.

Notably, for some pathways, such as “Hallmark_KRAS_Signaling_Down” which is positively enriched in both PDS-treated and control condition, we do not observe a significant higher overexpression in PDS cells (Figure 3-19). This means that its regulation is likely independent from PDS effect on MCF-7 cells.

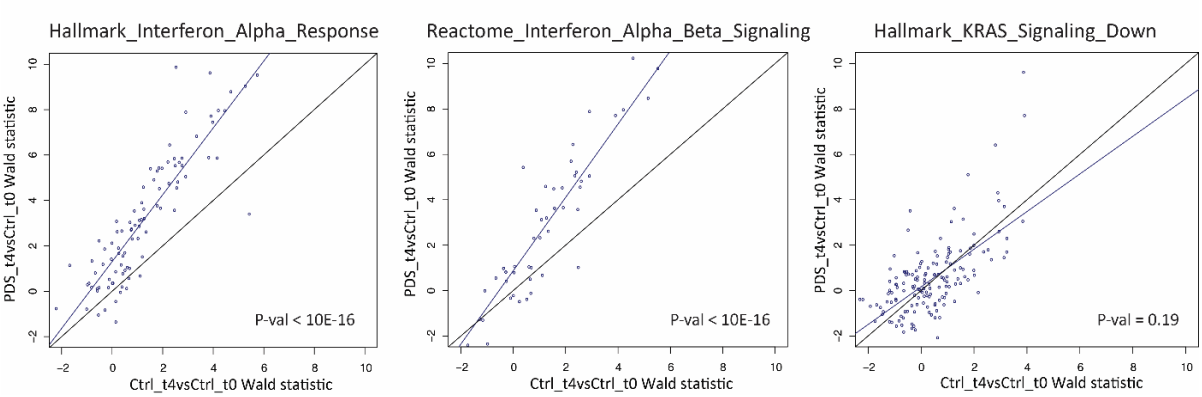

Figure 3-19 Scatterplots of Wald statistic for each gene of "Hallmark_interferon_alpha_response", “Reactome_Interferon_Alpha_Beta_Signaling” and “Hallmark_KRAS_signaling_down” gene sets. Each blue point represents a gene. Lines represent the bisector (black) and the regression line (blue). P-value of Kolmogorov-Smirnov test is reported in each scatterplot.

Furthermore, we looked at single gene level to investigate if genes encoding for chemokines, interleukins and interferon beta are likely more up regulated in PDS_t4 vs. Control_t0 than in Control_t4 vs. Control_t0 contrast. Using “cytokine-cytokine receptor interaction” KEGG pathway and Pathview R library, we observed that in many genes like IFNB1, chemokines like *CXCL1*, *CXCL2*, *CXCL3*, *CXCL10*, *CXCL11*, *CXCL12* and many others we have an up-regulation only in PDS treated cells and not in untreated condition (Figure 3-20).

Altogether, the data suggest that PDS can induce innate immune response in human cancer cells, via induction of micronuclei, which may then trigger the nucleic acid sensors (likely *cGAS*) and initiate the interferon stimulated response dependent on STING (unpublished data).

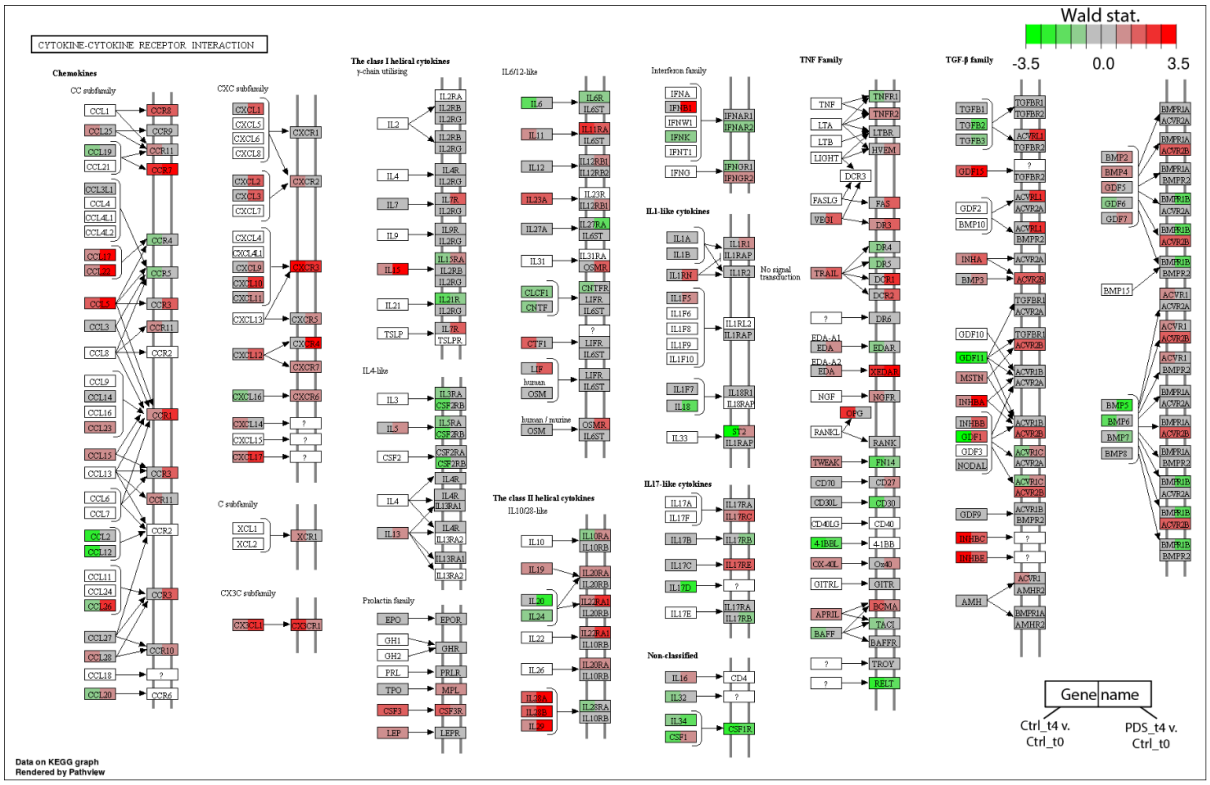

Figure 3-20 Schematic representation of cytokine-cytokine receptor interaction KEGG pathway. For each gene symbol is reported the Wald statistic of DEseq2 analysis for Ctrl_t4 vs. Ctrl_t0 (left) and PDS_t4 vs. Ctrl_t0 (right) contrasts. White boxes represent genes that are not expressed. Green (-3.5) to red (3.5) colour represent the Wald statistic value (as in legend).

3.2.2 Comparison of gene expression patterns between MCF-7 and U2OS cells.

We have also performed experiments to establish the gene expression profiles following PDS treatment in human sarcoma U2OS cells for a comparison with MCF-7 cells. In this case, we did two biological replicates and evaluated gene expression after 24 hours of PDS treatment.

Two biological replicates of RNA-seq libraries were prepared for the following samples and conditions:

- U2OS cells treated with PDS for 24 hours + 1 days of recovery in drug free medium (U2OS_PDS24)
- U2OS cells untreated for 24 hours+ 1 days of recovery in drug free medium (U2OS_Ctrl24)

Libraries followed the same analysis pipeline as the MCF-7 samples: they were processed by reads trimming with Trimmomatic, aligned on human genome (hg 19 version) using Hisat2 software and gene quantification was performed with Stringtie. As for MCF-7 samples, we have a high percentage of properly mapped reads (Table3-8).

Library	Library dimension (n. of reads)	Mapped reads (% of total)	Properly paired reads (% of total)
U2OS_Ctrl24 Rep.1	53,342,828	97.64	94.68
U2OS_Ctrl24 Rep.2	55,582,135	97.59	94.51
U2OS_PDS24 Rep.1	64,202,747	97.63	94.75
U2OS_PDS24 Rep.2	45,165,871	97.72	94.66

Table 3-8 Results of RNA-seq libraries alignment oh human genome(hg19).

Differential gene expression analysis with DEseq2 showed that 239 genes are significantly up-regulated in PDS-treated cells, while 100 genes are down-regulated.

In contrast to MCF-7 cells, GSEA results showed that while PDS-treated U2OS cells have a significantly enrichment in inflammatory response regulation, there is no significative enrichment in interferon response (Figure 3-21). As interferon response is mainly regulated by cGAS/STING pathway that is likely impaired in U2OS cells due to absence of STING, the absence of interferon response in U2OS may be due to the inactivity of the cGAS/STING pathway.

However, it is likely that in the absence of cGAS/STING pathway, other mechanisms of inflammatory response are induced by PDS in U2OS as inflammation response is unregulated (Figure 3-21).

Interestingly, most downregulated pathways are related to cell proliferation in both cells, although some pathways (e.g. oxidative phosphorylation and DNA repair) are downregulated in MCF-7 cells but not in U2OS, suggesting that PDS treatment may affect not only innate immune response pathways, but also other pathways in a specific way depending on cell lines.

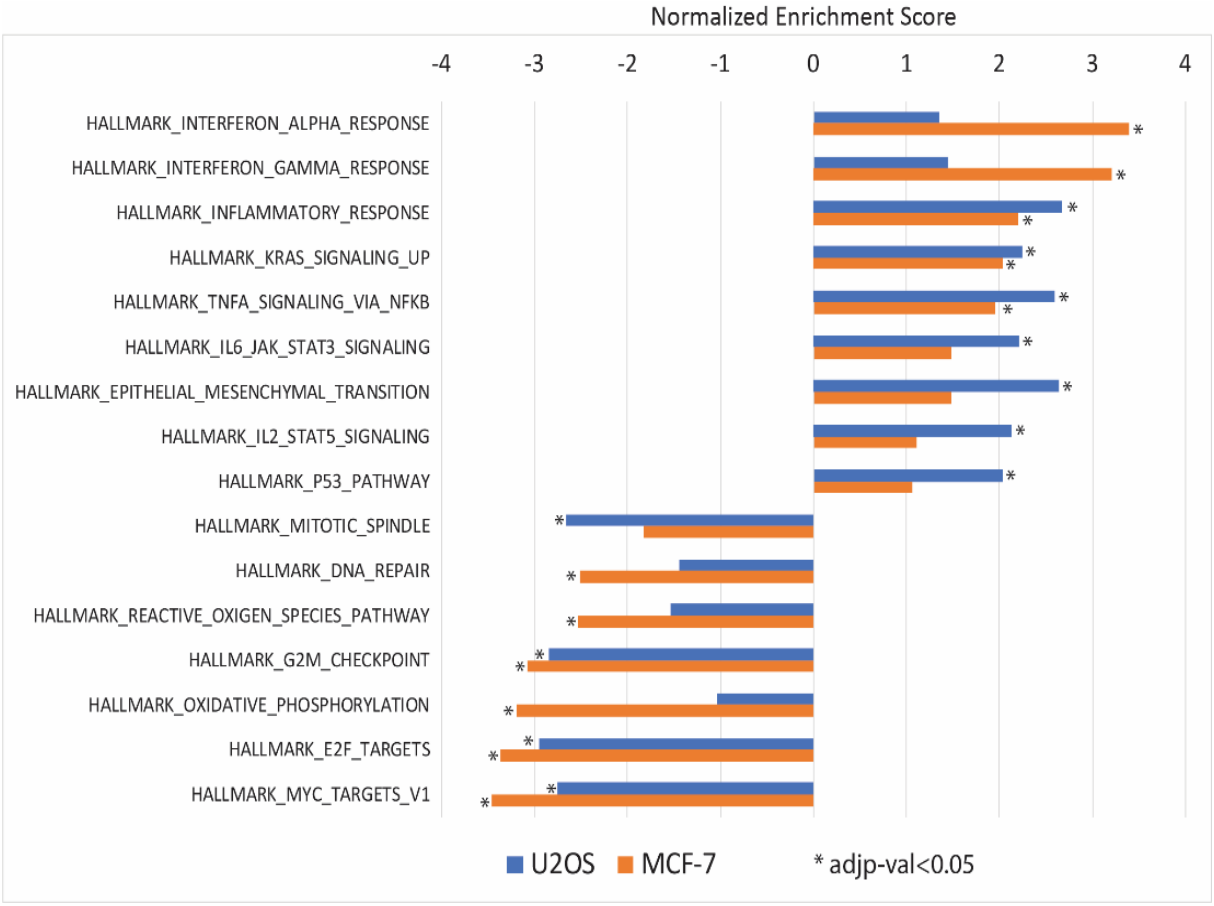

Figure 3-21 Bar plot with GSEA analysis results of PDS_t4 vs. Ctrl_t4 (in MCF7) and U2OS_PDS24 vs. U2OS_Ctrl24 contrasts using MsigDB hallmark gene sets. Only Top16 gene sets ranked by adjusted p-value are reported.

3.2.3 Innate immune response genes in human tumours: a PanCancer survey.

To evaluate the feasibility of using PDS as a potential immune response stimulator in cancer therapy, we aimed at understanding whether cGAS/STING pathway genes are mutated and how their expression is regulated in cancer tissues.

To do this, we analysed copy number variations (CNVs), mutations and gene expression data of four most important genes involved in cGAS/STING pathway (*cGAS*, *STING*, *TBK1* and *IRF3*) across 31 cancer types and ~7800 tumour samples from The Cancer Genome Atlas (TCGA) (Hoadley et al., 2018). The data used are part of the Network of Cancer Genes Database (Repana et al., 2019), maintained by the Ciccarelli group at The School of Cancer Studies of King's College London and The Francis Crick Institute.

The role of these four genes in tumour biology was evaluated in terms of:

i) mutation rates in each cancer type;
ii) gene expression alterations between cancer and tumour samples;
iii) survival rates of patients with different levels of cGAS/STING expression;
iv) correlation of cGAS/STING expression and immune-related tumour microenvironment.

To analyse mutation rate of cGAS/STING pathway genes, using copy number variations and mutations data, we distinguished 5 class of mutation based on the presence of copy number loss or amplification and of damaging mutations (defined as missense amino acid change as opposed to non-damaging mutations corresponding to nonsense mutations):

- Strong loss mutation: gene with a homozygous CNV loss or heterozygous CNV loss plus a damaging mutation
- Weak Loss mutation: gene with a heterozygous CNV loss or a damaging mutation
- Normal: absence of CNVs or mutations
- Weak Gain mutation: gene with a CNV amplification with copy number =3
- Strong Gain mutation: gene with a CNV amplification with copy number >3

Results of our analysis of mutations rates shows that CGAS/STING pathway genes are usually poorly mutated in cancer, with no homozygous loss-of-function mutations in almost

all cancer types. Most of the observed mutations are weak gene amplifications and heterozygous loss-of-function mutations (Figure 3-22).

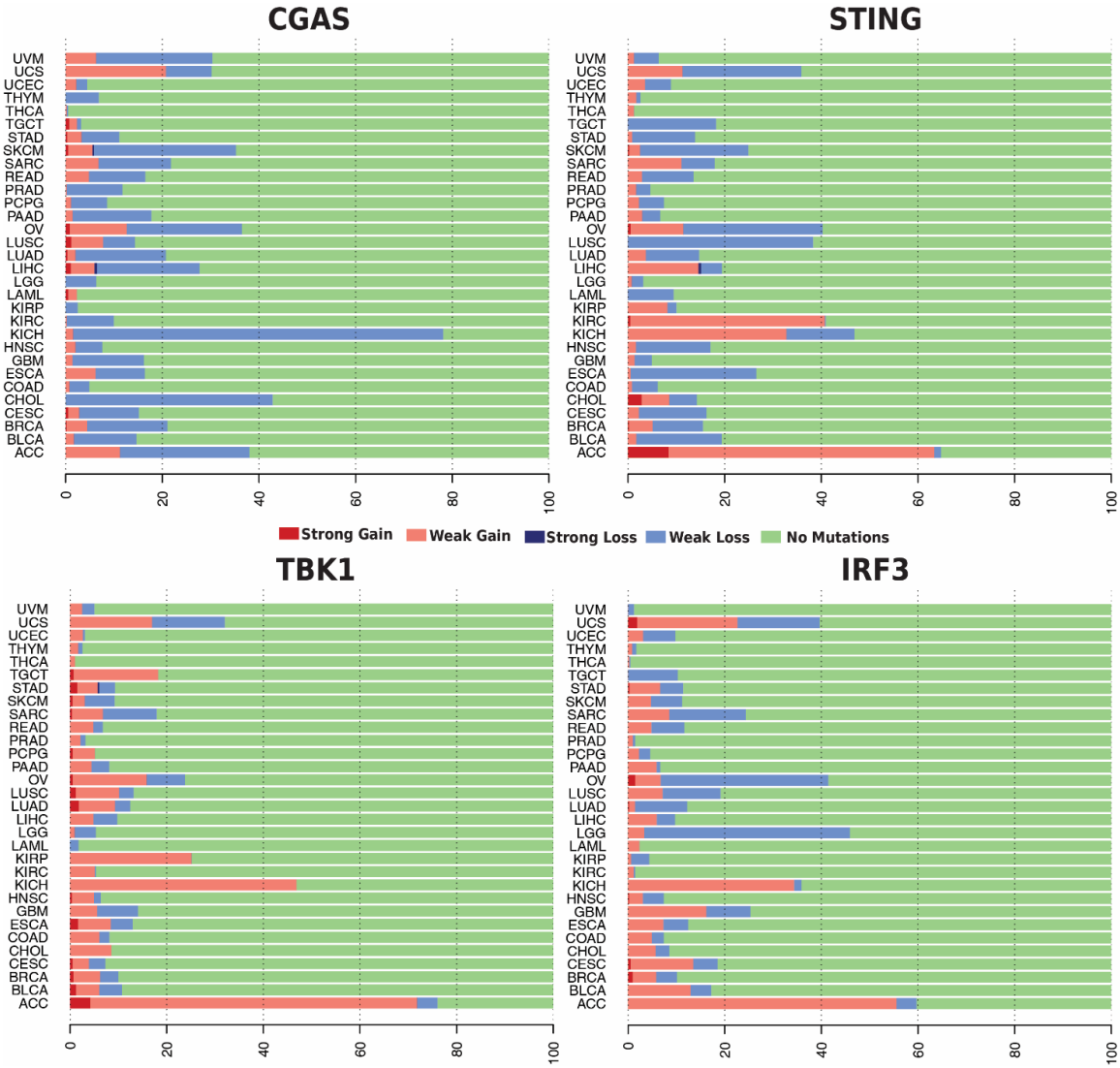

Figure 3-22 Proportion of samples for each cancer type with mutation in one of the CGAS/STING pathway genes. X-axis: proportion (%) of samples. Y-axis: TCGA cancer type.

Regarding gene expression of cGAS/STING pathway genes, we compared gene expression in cancer tissues matching these data with normal tissue ones (when present). We detected an altered expression of these genes in many cancer types compared to matched normal tissues (Figure 3-23). In particular, *cGAS* gene is significantly overexpressed in 11 cancer types out of the 22 for which we have expression data for both tumour and normal tissue. It shows a downregulation only in Kidney Chromophobe (KICH) and Prostate

adenocarcinoma (PRAD). *STING* gene is significantly overexpressed only in Kidney renal clear cell carcinoma (KIRC), Stomach adenocarcinoma (STAD) and Thyroid carcinoma (THCA), while it is underexpressed in Lung adenocarcinoma (LUAD) and Lung squamous cell carcinoma (LUSC). Two other key genes of this pathway, *TBK1* and *IRF3*, are overexpressed in 10 and 12 different cancer types, respectively. Notably, in KIRC, all the genes of the cGAS/STING pathways are overexpressed (Figure 3-23).

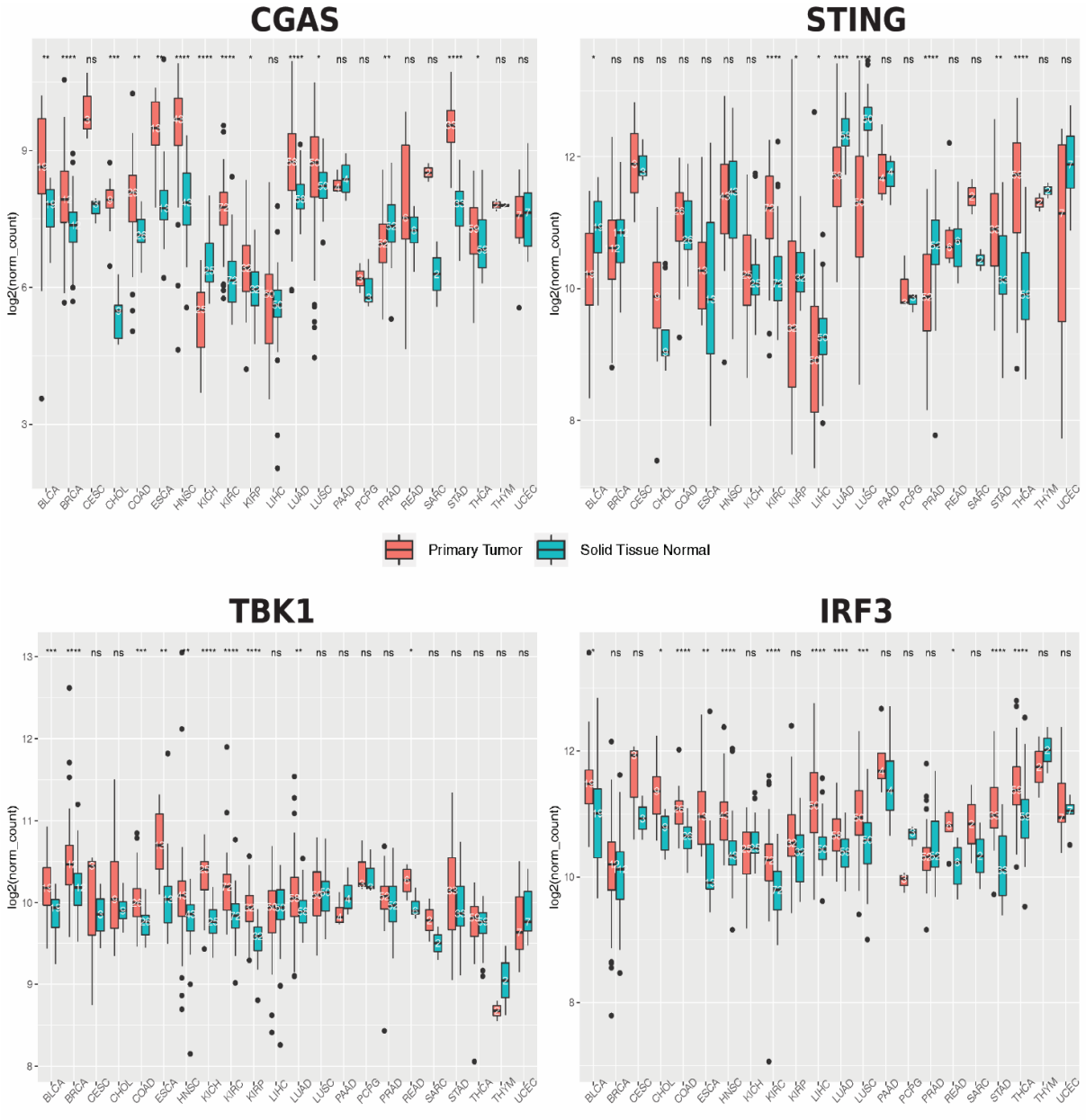

Figure 3-23 Expression of cGAS, STING, TBK1, IRF3 in primary tumour samples (red) compared to normal tissue samples (blue) for each cancer type. Symbols on top indicates p-value of Wilcoxon test (ns: p-

*val>0.05, *: pval<0.05, **: pval<0.01, ***: pval<0.001, ****: pval<0.0001). Values into boxplot indicate number of samples.*

Then, we tested if cGAS/STING pathway overexpression can be a predictor of survival. For each cancer type, we distinguished 2 group of samples for each gene: one "High", with the upper 33 percent of samples ranked by gene expression, and one "Low" with the lower 33 percent. Among all the cancer types, we obtained significant result only with *STING* and IRF3 (in only one cancer type) expression. Notably, *STING* overexpression leads to different outcome of survival, depending on cancer type (Figure 3-24). In particular, *STING* overexpression is a predictor of poor prognosis in Acute Myeloid Leukaemia (LAML), Uterine Corpus Endometrial Carcinoma (UCEC), Colon adenocarcinoma (COAD) and Rectum adenocarcinoma (READ), while in Brain Lower Grade Glioma (LGG), Skin Cutaneous Melanoma (SKCM) and THCA it is a predictor of better prognosis. Among the other cGAS/STING pathway genes, only *IRF3* overexpression predicts a significant poor prognosis in LAML, as for *STING*.

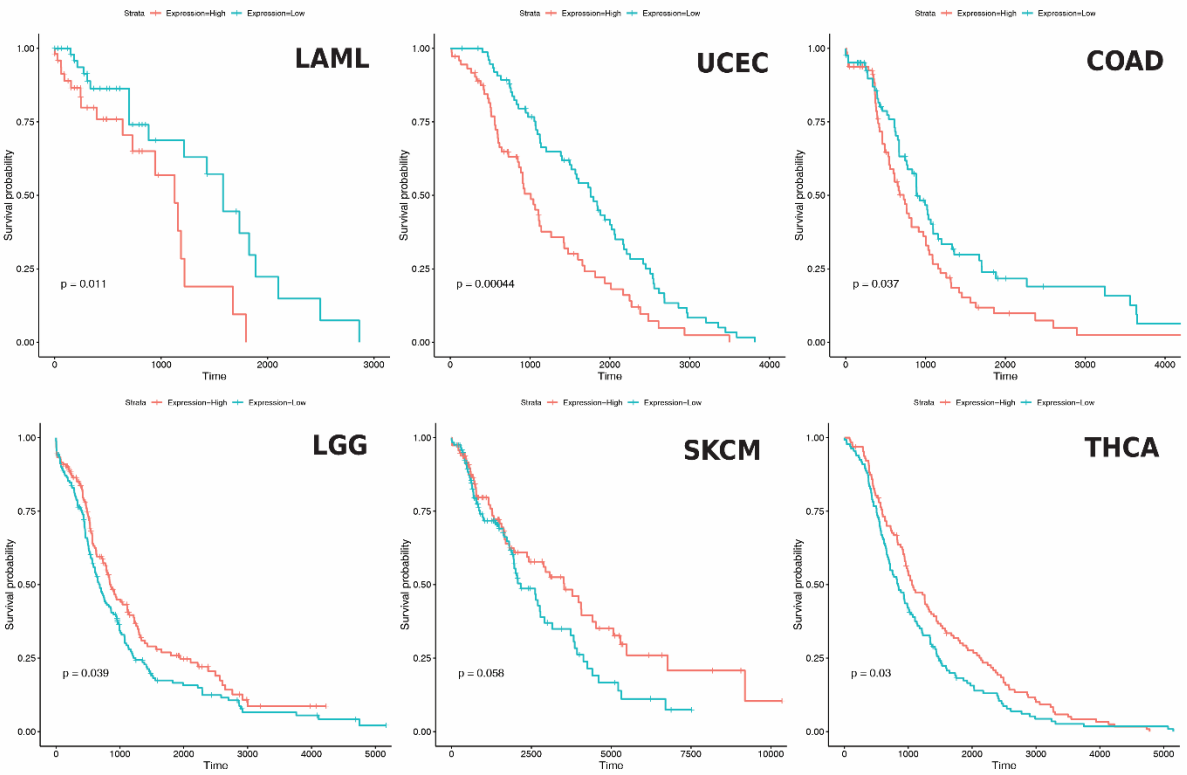

Figure 3-24 Overall Survival plot for STING expression in selected cancer types. Red line: High expression group (33th higher percentile of samples ranked by STING expression). Blue line: Low expression group (33th lower percentile of samples ranked by STING expression).

To elucidate the interplay of cGAS/STING pathway and tumour immune microenvironment, we correlated expression of this genes with leukocyte fraction, scores regarding signatures of immune cell infiltration in tumour samples and with gene set enrichment scores regarding the induction of innate immune response (Thorsson et al., 2018).

Results of this analysis show that *STING* gene expression is positively correlated with immune cells infiltration and interferon response in almost all cancer types, underlining its role in innate immune response activation in cancer (Figure 3-25). *CGAS* gene expression is well correlated in many cancer types, in particular in KICH and LGG cancer types (Figure 3-25). On the other hand, *TBK1* and *IRF3* gene expression is correlated with immune cells presence and interferon response only in few specific cancer types (Figure 3-25).

A possible explanation of this result is that *TBK1* and *IRF3* are usually expressed at a consistent level and that their activity is modulated by protein modifications. Notably, in LGG cancer type, *CGAS*, *STING* and *IRF3* expression are strongly correlated with immune infiltration, underlining the importance of this pathway in innate immune response activation in this cancer type.

Overall, the findings of the PanCancer survey showed that genes involved in cGAS/STING pathway are poorly mutated in most cancer types and that on the base of cancer type their expression can be sensibly altered if compared to normal tissue. Moreover, their expression positively correlates with immune infiltration and innate immune response activation, in particular for *STING* gene, the expression of which could be a predictor of survival in some cancer types. On the other hand, in some cases pathway overexpression is correlated with poor prognosis, suggesting that in some cancer types the activation of cGAS/STING might be deleterious.

Notably, in LGG (Low Grade Glioblastoma) there is a strong correlation between this pathway expression and innate immune response activation and *STING* expression is also a predictor of good prognosis, suggesting that modulation of this pathway in this cancer type can be a promising therapeutic strategy.

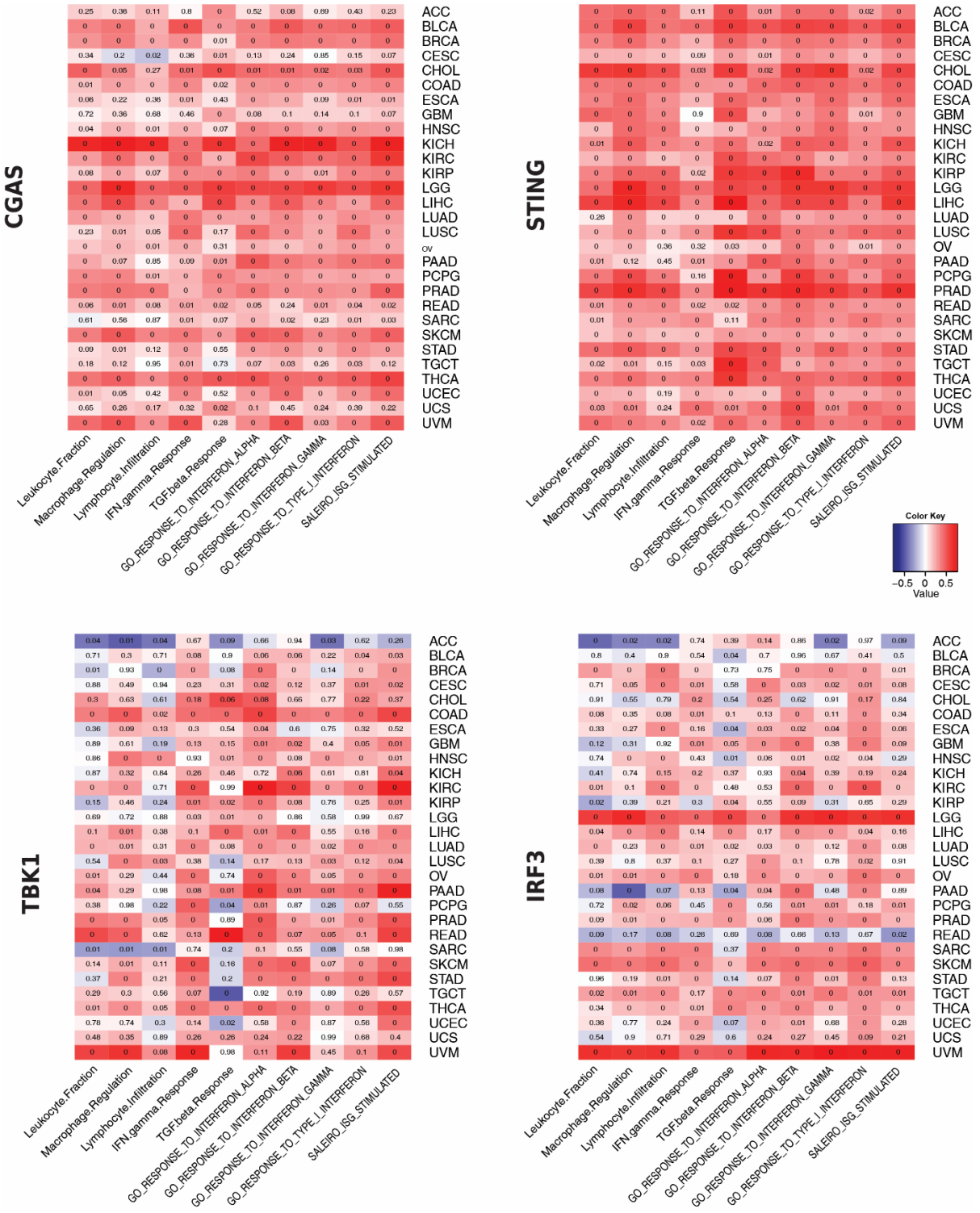

Figure 3-25 Heatmaps showing spearman correlation between gene expression of cGAS, STING, TBK1 and IRF3 and signature or enrichment score for each cancer type. Value of correlation is indicated in colour legend. Numbers within each cell indicate the p-value of correlation test.

4. DISCUSSION

The first part of my PhD project focused the definition of R-loop dynamics in cancer cells exposed to PDS and FG, two unrelated G4 binders. Through analysis of DRIP-sequencing and immunofluorescence microscopy data (De Magis et al. 2019; De Magis 2016), we observed that in U2OS cells there is an increase of R-loop structure after 5 minutes of exposure to two different G4 binders, FG and PDS (Table 3-4). These R-loops are mainly localized at promoter and transcription termination regions and are positively correlated with gene expression and the presence of CpG islands, in agreement with previous studies on R-loop mapping (Ginno et al., 2013, 2012). One of the major effects was that in many loci (1000 for FG and 639 for PDS) R-loop increase was due to an extension of pre-existent R-loops (Figure 3-10), prevalently at intronic regions and terminator regions (Figure 3-11). The data thus indicate that G4 binders induce unscheduled nuclear R-loops in U2OS cancer cells and that in many loci R-loop formation can be structurally compatible with binder stabilised G4 structures (section 3.1, Figure 4-1) (De Magis et al., 2019).

As R-loop regions identified by DRIP-sequencing lack strand information, which is essential to better define the structural relation of R-loops with G4, annotation of R-loop to the correct gene became a crucial step in my analysis. I spent part of my PhD project on the development of a tool that can perform a better annotation of R-loop peaks comparing to other existing annotation tool. These efforts led to the release of a new annotation tool, DROPA (DRIP Optimized Peak Annotator) (Russo et al., 2019). The main improvement comparing to other broadly used tools relies on the possibility to annotate peaks on the base of gene expression. Since R-loops are mainly co-transcriptional, when DRIP-seq peaks overlap more than one gene, DROPA assigns the peak to the expressed (or most expressed) gene. DROPA was evaluated in terms of efficiency in annotating peaks to the correct gene using stranded R-loop datasets (obtained via DRIPc-seq technique), resulting in 93.8% of correct assignment (see section 3.1.2). I also performed a comparison with other broadly used tools, showing that DROPA performs a more robust and reliable annotation. Even though DROPA showed a good general efficiency in R-loop annotation, there are some conditions that can limit the proper assignment of R-loop peaks to the correct gene. The main one regards

antisense R-loop peaks (Sanz et al., 2016) that cannot be detected by DROPA. To compensate for this limitation, DROPA provides a list of peaks assigned only to the upstream/downstream region of expressed gene, where antisense transcripts can be present. A comparison of this information with genomic datasets of antisense transcripts can help the user for further analysis. Another limitation of DROPA regards the fact that in some cases RNA-seq methodology may not represent the nascent transcript levels level in the cell, leading to false positive annotation. This limitation can be overcome using, instead of RNA-seq gene expression data, nascent RNA data that can be obtained using GRO seq (Core et al., 2008), PRO-seq (Kwak et al., 2013) or TT-seq (Schwalb et al., 2016) methodologies.

Thus, taking advantage of DROPA, we predicted strandness of DRIP-seq peaks and found that extended regions of R-loops were significantly enriched in G4 structures when the latter ones form on the displaced strand of the R-loop. These results are in agreement with Duquette et al. 2004 that showed a structural compatibility of R-loops with G-quadruplexes structures *in vitro* and in *E.coli*. These findings let us to propose a model in which the treatment of cancer cells with G4-binders leads to stabilization of G4s in open transcribed chromatin region, in agreement with other studies (Hänsel-Hertsch et al., 2016), and consequently R-loop structures are also stabilized, allowing the extension of DNA:RNA hybrids (Figure 4-1).

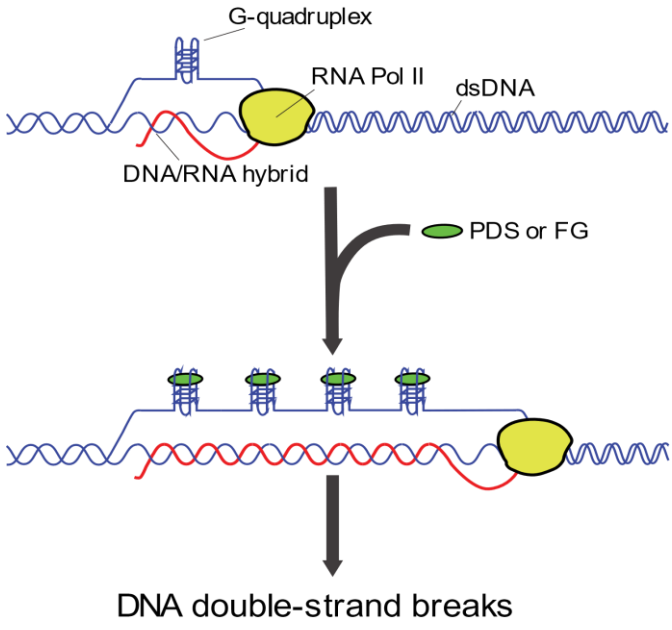

Figure 4-1 Model of G4-binders activity in U2OS cells. Adapted from De Magis et al., 2019.

As presented in the introduction, a mechanistic link between unscheduled R-loop formation and DNA damage and genome instability has been shown in different studies (Crossley et al., 2019). We also found that PDS and FG can induce DNA double-strand breaks and DNA damage response (DDR) after few hours of exposure to the tested G4 binders (Figure 4-1). As DNA damage and DDR were abolished by overexpressing *RNaseH1*, our data strongly support that G4 binders promote DNA cleavage and DDR by increasing the levels of nuclear R loops (De Magis et al., 2019).

In addition, we also found that PDS and FG showed different effect on U2OS cells: while FG showed a more cytotoxic effect, PDS was much less cytotoxic and induced micronuclei formation in a R-loop-dependent way (De Magis et al., 2019). We still do not know the mechanistic basis of this difference. However, we can speculate that it may be partially due to different G4 binding specificity of the two compounds, leading to stabilization of R-loop in different loci and consequently leading to diverse downstream effects.

Since micronuclei, a well-known marker of genome instability (Hatch et al., 2013), were recently reported to be a source of cytosolic DNA and stimulate the cGAS/STING pathway and an innate immune response (Harding et al., 2017; MacKenzie et al., 2017), in the second part of my PhD project I assessed whether PDS can be an effective immune stimulator compound. In such case, G4 binders might be proposed in clinical settings as coadjutant agent in cancer immunotherapy instead of as cytotoxic compounds. As STING pathway has been shown to be impaired in U2OS cells (Deschamps and Kalamvoki, 2017), we decided to test PDS effects on MCF-7 breast and U2OS sarcoma cancer cell lines by determining gene expression profiles by Illumina RNA-seq analyses after G4 binder treatment. In MCF-7 cells treated with PDS, we observed a strong and clear increase of micronuclei formation (data not shown). Furthermore, through immunofluorescence experiments, we observed that micronuclei are recognized by cGAS, in agreement with previous studies (Harding et al., 2017; MacKenzie et al., 2017) (data not shown). Under similar conditions, RNA-seq data showed a significant upregulation of genes related to inflammation and type I interferon response.

Since in our experimental design we included both a control condition after 4 days of recovery (as PDS treated cells) and one other control condition at time 0, we were able to

detect changes in gene expression that are independent from PDS. Interestingly, in untreated cells, after 4 days of recovery we observed a weaker but significant enrichment in type I interferon response (Fig 3-17). We then demonstrated that in PDS treated cells expression of genes belonging to interferon alpha response dataset are significantly more upregulated than control_t4 cells (Fig 3-18). Thus, as immunofluorescence microscopy data showed that micronuclei are also increased in untreated cells after 4-5 days in culture, although to a lower extent than in PDS-treated cells, the findings show that MCF-7 cells are likely characterized by genome instability leading to micronuclei formation and that treatment with PDS exacerbates the process and leads to a stronger type I interferon response. These evidences were validated via qPCR experiment on some representative genes (*IFNB1*, *CXCL10*, *CCL5*, *IFIT1*, *IFI44*, *DDX60*) and we observed that when STING activity is inhibited (using specific siRNAs or a chemical inhibitor) induction of these gene is abolished (data not shown).

We then performed a comparison of the effects of PDS treatment on gene expression between MCF-7 and U2OS cells. Temporal sets were different between the cell lines as U2OS cells were collected after 1 day of recovery, while MCF-7 cells after 4 days. U2OS RNA-seq experiment was set mainly to evaluate the putative effects of G4 binders on transcription and has not been reported in our previous publication (De Magis et al., 2019). However, a comparison between U2OS and MCF-7 lines can provide some information on PDS biological effects.

Interestingly, we observed marked differences between U2OS and MCF-7 cells in relation to gene expression profiles upon PDS treatment and micronuclei increase. First, no enrichment of type I interferon response genes was detected in U2OS cells (Figure 3-21). This finding may be due to the absence of *STING* expression as reported by others (Deschamps and Kalamvoki, 2017). Interestingly, U2OS cells showed an upregulated inflammatory response may be mediated by activation of alternative immune pathway (IL6/JAK/STAT3 and IL2/STAT5 signalling). When considering the down regulated gene pathways, both U2OS and MCF-7 have a similar response in terms of cell proliferation inhibition and cell cycle arrest. The findings thus suggest that a properly working STING pathway is required to trigger a type I interferon response to PDS treatment in cancer cells. Nevertheless, other immune-related pathways

appear to be activated in cells treated with PDS that are independent from STING. These pathways remain to be characterised fully.

In the last part of my PhD project, I investigated how cGAS/STING pathway genes are altered in human cancer tissues, in collaboration with Prof. Francesca Ciccarelli. By using TCGA data regarding 31 different cancer types, I showed that while cGAS/STING pathway genes are generally poorly mutated in human cancers, gene expression can be substantially altered in many cancer types. Except for *STING* gene that is downregulated in lung adenocarcinoma and lung squamous cell carcinoma (Figure 3-23), *cGAS*, *TBK1* and *IRF3* are upregulated in many cancer types. Furthermore, we observed that expression is correlated with immune cells infiltration and interferon response in most of cancer types. The highest correlation was found in case of STING expression and, to lower extent, for *cGAS* expression, suggesting that these two genes may be the main actor in innate immune response in human cancers.

However, we found that cGAS/STING higher expression poorly correlate with better prognosis. In fact, only in Brain Lower Grade Glioma (LGG), Skin Cutaneous Melanoma (SKCM) and Thyroid carcinoma (THCA) STING overexpression is predictor of better prognosis, while in some cancer types (Acute Myeloid Leukaemia, Uterine Corpus Endometrial Carcinoma, Colon adenocarcinoma and Rectum adenocarcinoma) *STING* is a predictor of poor prognosis (Figure 3-24). As cGAS/STING pathway is reported as tumour suppressor, these findings may be surprising at first sight. However, recently it has been reported that a chronic activation of STING pathway due to accumulation of cytosolic DNA may lead to immune evasion of tumour, therapeutic resistance and metastasis development (Bakhoum et al. 2018). These findings, together with what we reported on STING pathway in PanCancer, suggest that the pathway may be a double-edged sword. When STING pathway is transiently stimulated by cytosolic DNA (e.g. via induction of micronuclei with specific treatments) type I interferon stimulation leads to a proinflammatory response, T-cell recruitment to tumour microenvironment and cancer cell senescence. On the other hand, when cancer cells accumulate DNA damage and persistent genome instability, they show a chronic activation of STING pathway, and interferon signalling is progressively downregulated leading to immune evasion and metastasis (Bakhoum and Cantley 2018). Therefore, STING pathway modulation may be a suitable therapeutic strategy for specific cancer patients only.

In conclusion, my project involved both bioinformatic analysis of genomic and transcriptomic data and development of a new in-silico tool to address a specific biological question. The first part of my project, together with the efforts of my collaborators, leads to the discovery of a mechanistic model of DNA damage in cancer cells mediated by R-loop stabilization after stabilization of G4 structures in the displaced strand using different G4-binders (De Magis et al. 2019). Moreover, I developed a new peak annotation tool to annotate DRIP-seq peaks which was crucial in our work to investigate R-loop/G4s interplay and may be useful to scientific community in the analysis of DRIP-seq data (Russo et al. 2019).

In the second part of my PhD project we investigated the feasibility of pyridostatin as potential anticancer drug able to elicit innate immune response in cancer cells. Our findings suggest that in cells in which cytosolic DNA sensor pathway (cGAS/STING) is functional (MCF-7 cells), there is an effective stimulation of interferon response. Notably, in case of STING pathway impairment (in U2OS cells) interferon response may be replaced by other inflammatory signalling, mediated by interleukins (these results are part of a further publication, in preparation). Our data support a further investigation of pyridostatin ability to stimulate a proinflammatory tumour environment and to trigger T-cell recruitment in cancer tissues. However, as the stimulation of the cGAS/STING pathway may lead to opposite outcomes, depending on cancer type and stage, the data already indicate that a potential anticancer strategy using PDS as immune stimulator may be effective in certain cancer types but not in others depending on tumour microenvironment and existing immune activation.

5. BIBLIOGRAPHY

Allhoff, M., Seré, K., Chauvistré, H., Lin, Q., Zenke, M., Costa, I.G., 2014. Detecting differential peaks in ChIP-seq signals with ODIN. Bioinformatics 30, 3467–3475. https://doi.org/10.1093/bioinformatics/btu722

Amato, J., Morigi, R., Pagano, B., Pagano, A., Ohnmacht, S., De Magis, A., Tiang, Y.P., Capranico, G., Locatelli, A., Graziadio, A., Leoni, A., Rambaldi, M., Novellino, E., Neidle, S., Randazzo, A., 2016. Toward the Development of Specific G-Quadruplex Binders: Synthesis, Biophysical, and Biological Studies of New Hydrazone Derivatives. J. Med. Chem. 59, 5706–5720. https://doi.org/10.1021/acs.jmedchem.6b00129

Bartsch, K., Knittler, K., Borowski, C., Rudnik, S., Damme, M., Aden, K., Spehlmann, M.E., Frey, N., Saftig, P., Chalaris, A., Rabe, B., 2017. Absence of RNase H2 triggers generation of immunogenic micronuclei removed by autophagy. Hum. Mol. Genet. 26, 3960–3972. https://doi.org/10.1093/hmg/ddx283

Becherel, O.J., Sun, J., Yeo, A.J., Nayler, S., Fogel, B.L., Gao, F., Coppola, G., Criscuolo, C., De Michele, G., Wolvetang, E., Lavin, M.F., 2015. A new model to study neurodegeneration in ataxia oculomotor apraxia type 2. Hum. Mol. Genet. 24, 5759–5774. https://doi.org/10.1093/hmg/ddv296

Bedrat, A., Lacroix, L., Mergny, J.L., 2016. Re-evaluation of G-quadruplex propensity with G4Hunter. Nucleic Acids Res. 44, 1746–1759. https://doi.org/10.1093/nar/gkw006

Biffi, G., Tannahill, D., McCafferty, J., Balasubramanian, S., 2013. Quantitative visualization of DNA G-quadruplex structures in human cells. Nat. Chem. 5, 182–186. https://doi.org/10.1038/nchem.1548

Boguslawski, S.J., Smith, D.E., Michalak, M.A., Mickelson, K.E., Yehle, C.O., Patterson, W.L., Carrico, R.J., 1986. Characterization of monoclonal antibody to DNA · RNA and its application to immunodetection of hybrids. J. Immunol. Methods 89, 123–130. https://doi.org/10.1016/0022-1759(86)90040-2

Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120.

https://doi.org/10.1093/bioinformatics/btu170

Bonnet, A., Grosso, A.R., Elkaoutari, A., Coleno, E., Presle, A., Sridhara, S.C., Janbon, G., Géli, V., de Almeida, S.F., Palancade, B., 2017. Introns Protect Eukaryotic Genomes from Transcription-Associated Genetic Instability. Mol. Cell 67, 608-621.e6. https://doi.org/10.1016/j.molcel.2017.07.002

Brosh, R.M., Cantor, S.B., 2014. Molecular and cellular functions of the FANCJ DNA helicase defective in cancer and in Fanconi anemia. Front. Genet. https://doi.org/10.3389/fgene.2014.00372

Chambers, V.S., Marsico, G., Boutell, J.M., Di Antonio, M., Smith, G.P., Balasubramanian, S., 2015. High-throughput sequencing of DNA G-quadruplex structures in the human genome. Nat. Biotechnol. 33, 877–881. https://doi.org/10.1038/nbt.3295

Chang, E.Y.C., Novoa, C.A., Aristizabal, M.J., Coulombe, Y., Segovia, R., Chaturvedi, R., Shen, Y., Keong, C., Tam, A.S., Jones, S.J.M., Masson, J.Y., Kobor, M.S., Stirling, P.C., 2017. RECQ-like helicases Sgs1 and BLM regulate R-loop- associated genome instabil. J. Cell Biol. 216, 3991–4005. https://doi.org/10.1083/jcb.201703168

Chen, L., Chen, J.Y., Zhang, X., Gu, Y., Xiao, R., Shao, C., Tang, P., Qian, H., Luo, D., Li, H., Zhou, Y., Zhang, D.E., Fu, X.D., 2017. R-ChIP Using Inactive RNase H Reveals Dynamic Coupling of R-loops with Transcriptional Pausing at Gene Promoters. Mol. Cell 68, 745-757.e5. https://doi.org/10.1016/j.molcel.2017.10.008

Chen, P.B., Chen, H. V., Acharya, D., Rando, O.J., Fazzio, T.G., 2015. R loops regulate promoter-proximal chromatin architecture and cellular differentiation. Nat. Struct. Mol. Biol. 22, 999–1007. https://doi.org/10.1038/nsmb.3122

Core, L.J., Waterfall, J.J., Lis, J.T., 2008. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science (80-.). 322, 1845–1848. https://doi.org/10.1126/science.1162228

Crabbe, L., Verdun, R.E., Haggblom, C.I., Karlseder, J., 2004. Defective telomere lagging strand synthesis in cells lacking WRN helicase activity. Science (80-.). 306, 1951–1953. https://doi.org/10.1126/science.1103619

Cristini, A., Ricci, G., Britton, S., Salimbeni, S., Huang, S.-Y.N., Marinello, J., Calsou, P., Pommier, Y., Favre, G., Capranico, G., Gromak, N., Sordet, O., 2019. Dual Processing of R-Loops and Topoisomerase I Induces Transcription-Dependent DNA Double-Strand Breaks. Cell Rep. 28, 3167-3181.e6. https://doi.org/10.1016/j.celrep.2019.08.041

Crossley, M.P., Bocek, M., Cimprich, K.A., 2019. R-Loops as Cellular Regulators and Genomic Threats. Mol. Cell 73, 398–411. https://doi.org/10.1016/j.molcel.2019.01.024

De Magis, A., Manzo, S.G., Russo, M., Marinello, J., Morigi, R., Sordet, O., Capranico, G., 2019. 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. https://doi.org/10.1073/pnas.1810409116

Deschamps, T., Kalamvoki, M., 2017. Impaired STING Pathway in Human Osteosarcoma U2OS Cells Contributes to the Growth of ICP0-Null Mutant Herpes Simplex Virus. J. Virol. 91, e00006-17. https://doi.org/10.1128/jvi.00006-17

Domínguez-Sánchez, M.S., Barroso, S., Gómez-González, B., Luna, R., Aguilera, A., 2011. Genome instability and transcription elongation impairment in human cells depleted of THO/TREX. PLoS Genet. 7. https://doi.org/10.1371/journal.pgen.1002386

Drolet, M., Bi, X., Liu, L.F., 1994. Hypernegative supercoiling of the DNA template during transcription elongation in vitro. J. Biol. Chem. 269, 2068–2074.

Drosopoulos, W.C., Kosiyatrakul, S.T., Schildkraut, C.L., 2015. BLM helicase facilitates telomere replication during leading strand synthesis of telomeres. J. Cell Biol. 210, 191–208. https://doi.org/10.1083/jcb.201410061

Dumelie, J.G., Jaffrey, S.R., 2017. Defining the location of promoter-associated R-loops at near-nucleotide resolution using bisDRIP-seq. Elife 6. https://doi.org/10.7554/elife.28306

Duquette, M.L., Handa, P., Vincent, J.A., Taylor, A.F., Maizels, N., 2004. Intracellular transcription of G-rich DNAs induces formation of G-loops, novel structures containing G4 DNA. Genes Dev. 18, 1618–1629. https://doi.org/10.1101/gad.1200804

El Hage, A., Webb, S., Kerr, A., Tollervey, D., 2014. Genome-Wide Distribution of RNA-DNA Hybrids Identifies RNase H Targets in tRNA Genes, Retrotransposons and Mitochondria.

PLoS Genet. 10, e1004716. https://doi.org/10.1371/journal.pgen.1004716

Gan, W., Guan, Z., Liu, J., Gui, T., Shen, K., Manley, J.L., Li, X., 2011. R-loop-mediated genomic instability is caused by impairment of replication fork progression. Genes Dev. 25, 2041–2056. https://doi.org/10.1101/gad.17010011

GELLERT, M., LIPSETT, M.N., DAVIES, D.R., 1962. Helix formation by guanylic acid. Proc. Natl. Acad. Sci. U. S. A. 48, 2013–2018. https://doi.org/10.1073/pnas.48.12.2013

Ginno, P.A., Lim, Y.W., Lott, P.L., Korf, I., Chédin, F., 2013. GC skew at the 59 and 39 ends of human genes links R-loop formation to epigenetic regulation and transcription termination. Genome Res. 23, 1590–1600. https://doi.org/10.1101/gr.158436.113

Ginno, P.A., Lott, P.L., Christensen, H.C., Korf, I., Chédin, F., 2012. R-Loop Formation Is a Distinctive Characteristic of Unmethylated Human CpG Island Promoters. Mol. Cell 45, 814–825. https://doi.org/10.1016/j.molcel.2012.01.017

Gold, B., Cankovic, M., Furtado, L. V., Meier, F., Gocke, C.D., 2015. Do Circulating Tumor Cells, Exosomes, and Circulating Tumor Nucleic Acids Have Clinical Utility? J. Mol. Diagnostics 17, 209–224. https://doi.org/10.1016/j.jmoldx.2015.02.001

Gratia, M., Rodero, M.P., Conrad, C., Bou Samra, E., Maurin, M., Rice, G.I., Duffy, D., Revy, P., Petit, F., Dale, R.C., Crow, Y.J., Amor-Gueret, M., Manel, N., 2019. Bloom syndrome protein restrains innate immune sensing of micronuclei by cGAS. J. Exp. Med. jem.20181329. https://doi.org/10.1084/jem.20181329

Gray, L.T., Vallur, A.C., Eddy, J., Maizels, N., 2014. G quadruplexes are genomewide targets of transcriptional helicases XPB and XPD. Nat. Chem. Biol. 10, 313–318. https://doi.org/10.1038/nchembio.1475

Grunseich, C., Wang, I.X., Watts, J.A., Burdick, J.T., Guber, R.D., Zhu, Z., Bruzel, A., Lanman, T., Chen, K., Schindler, A.B., Edwards, N., Ray-Chaudhury, A., Yao, J., Lehky, T., Piszczek, G., Crain, B., Fischbeck, K.H., Cheung, V.G., 2018. Senataxin Mutation Reveals How R-Loops Promote Transcription by Blocking DNA Methylation at Gene Promoters. Mol. Cell 69, 426-437.e7. https://doi.org/10.1016/j.molcel.2017.12.030

Halász, L., Karányi, Z., Boros-Oláh, B., Kuik-Rózsa, T., Sipos, É., Nagy, É., Mosolygó-L, Á., Mázló,

A., Rajnavölgyi, É., Halmos, G., Székvölgyi, L., 2017. RNA-DNA hybrid (R-loop) immunoprecipitation mapping: An analytical workflow to evaluate inherent biases. Genome Res. 27, 1063–1073. https://doi.org/10.1101/gr.219394.116

Hamperl, S., Bocek, M.J., Saldivar, J.C., Swigut, T., Cimprich, K.A., 2017. Transcription-Replication Conflict Orientation Modulates R-Loop Levels and Activates Distinct DNA Damage Responses. Cell 170, 774-786.e19. https://doi.org/10.1016/j.cell.2017.07.043

Hänsel-Hertsch, R., Beraldi, D., Lensing, S. V., Marsico, G., Zyner, K., Parry, A., Di Antonio, M., Pike, J., Kimura, H., Narita, M., Tannahill, D., Balasubramanian, S., 2016. G-quadruplex structures mark human regulatory chromatin. Nat. Genet. 48, 1267–1272. https://doi.org/10.1038/ng.3662

Hänsel-Hertsch, R., Di Antonio, M., Balasubramanian, S., 2017. DNA G-quadruplexes in the human genome: Detection, functions and therapeutic potential. Nat. Rev. Mol. Cell Biol. 18, 279–284. https://doi.org/10.1038/nrm.2017.3

Hänzelmann, S., Castelo, R., Guinney, J., 2013. GSVA: Gene set variation analysis for microarray and RNA-Seq data. BMC Bioinformatics 14. https://doi.org/10.1186/1471-2105-14-7

Harding, S.M., Benci, J.L., Irianto, J., Discher, D.E., Minn, A.J., Greenberg, R.A., 2017. Mitotic progression following DNA damage enables pattern recognition within micronuclei. Nature 548, 466–470. https://doi.org/10.1038/nature23470

Hatch, E.M., Fischer, A.H., Deerinck, T.J., Hetzer, M.W., 2013. Catastrophic Nuclear Envelope Collapse in Cancer Cell Micronuclei. Cell 154, 47. https://doi.org/10.1016/j.cell.2013.06.007

Heijink, A.M., Talens, F., Jae, L.T., van Gijn, S.E., Fehrmann, R.S.N., Brummelkamp, T.R., van Vugt, M.A.T.M., 2019. BRCA2 deficiency instigates cGAS-mediated inflammatory signaling and confers sensitivity to tumor necrosis factor-alpha-mediated cytotoxicity. Nat. Commun. 10, 100. https://doi.org/10.1038/s41467-018-07927-y

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., Glass, C.K., 2010. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol. Cell

38, 576–89. https://doi.org/10.1016/j.molcel.2010.05.004

Hoadley, K.A., Yau, C., Hinoue, T., Wolf, D.M., Lazar, A.J., Drill, E.,[…], Ley, T., Van Tine, B., Westervelt, P., Rubin, M.A., Lee, J. Il, Aredes, N.D., Mariamidze, A., Stuart, J.M., Benz, C.C., Laird, P.W., 2018. Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. Cell 173, 291-304.e6. https://doi.org/10.1016/j.cell.2018.03.022

Hollingsworth, R.E., Jansen, K., 2019. Turning the corner on therapeutic cancer vaccines. npj Vaccines. https://doi.org/10.1038/s41541-019-0103-y

Huang, W., Loganantharaj, R., Schroeder, B., Fargo, D., Li, L., 2013. PAVIS: A tool for Peak Annotation and Visualization. Bioinformatics 29, 3097–3099. https://doi.org/10.1093/bioinformatics/btt520

Hunter, J.D., 2007. Matplotlib: A 2D graphics environment. Comput. Sci. Eng. 9, 99–104. https://doi.org/10.1109/MCSE.2007.55

Huppert, J.L., Balasubramanian, S., 2007. G-quadruplexes in promoters throughout the human genome. Nucleic Acids Res. 35, 406–413. https://doi.org/10.1093/nar/gkl1057

Huppert, J.L., Balasubramanian, S., 2005. Prevalence of quadruplexes in the human genome. Nucleic Acids Res. 33, 2908–2916. https://doi.org/10.1093/nar/gki609

Kalos, M., Levine, B.L., Porter, D.L., Katz, S., Grupp, S.A., Bagg, A., June, C.H., 2011. T cells with chimeric antigen receptors have potent antitumor effects and can establish memory in patients with advanced leukemia. Sci. Transl. Med. 3. https://doi.org/10.1126/scitranslmed.3002842

Kim, D., Paggi, J.M., Park, C., Bennett, C., Salzberg, S.L., 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat. Biotechnol. 37, 907–915. https://doi.org/10.1038/s41587-019-0201-4

Kondili, M., Fust, A., Preussner, J., Kuenne, C., Braun, T., Looso, M., 2017. UROPA: A tool for Universal RObust Peak Annotation. Sci. Rep. 7, 2593. https://doi.org/10.1038/s41598-017-02464-y

Kouzine, F., Wojtowicz, D., Baranello, L., Yamane, A., Nelson, S., Resch, W., Kieffer-Kwon, K.R.,

Benham, C.J., Casellas, R., Przytycka, T.M., Levens, D., 2017. Permanganate/S1 Nuclease Footprinting Reveals Non-B DNA Structures with Regulatory Potential across a Mammalian Genome. Cell Syst. 4, 344-356.e7. https://doi.org/10.1016/j.cels.2017.01.013

Kwak, H., Fuda, N.J., Core, L.J., Lis, J.T., 2013. Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. Science (80-.). 339, 950–953. https://doi.org/10.1126/science.1229386

Kwok, C.K., Merrick, C.J., 2017. G-Quadruplexes: Prediction, Characterization, and Biological Application. Trends Biotechnol. 35, 997–1013. https://doi.org/10.1016/j.tibtech.2017.06.012

Lan, Y.Y., Londoño, D., Bouley, R., Rooney, M.S., Hacohen, N., 2014. Dnase2a deficiency uncovers lysosomal clearance of damaged nuclear DNA via autophagy. Cell Rep. 9, 180–192. https://doi.org/10.1016/j.celrep.2014.08.074

Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R., Pfister, H., 2014. UpSet: Visualization of intersecting sets. IEEE Trans. Vis. Comput. Graph. 20, 1983–1992. https://doi.org/10.1109/TVCG.2014.2346248

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Lim, Y.W., Sanz, L.A., Xu, X., Hartono, S.R., Chédin, F., 2015. Genome-wide DNA hypomethylation and RNA:DNA hybrid accumulation in Aicardi–Goutières syndrome. Elife 4, e08007. https://doi.org/10.7554/elife.08007

Love, M.I., Huber, W., Anders, S., 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15. https://doi.org/10.1186/s13059-014-0550-8

Luo, W., Brouwer, C., 2013. Pathview: An R/Bioconductor package for pathway-based data integration and visualization. Bioinformatics 29, 1830–1831. https://doi.org/10.1093/bioinformatics/btt285

MacKenzie, K.J., Carroll, P., Martin, C.A., Murina, O., Fluteau, A., Simpson, D.J., Olova, N., Sutcliffe, H., Rainger, J.K., Leitch, A., Osborn, R.T., Wheeler, A.P., Nowotny, M., Gilbert, N., Chandra, T., Reijns, M.A.M., Jackson, A.P., 2017. CGAS surveillance of micronuclei links genome instability to innate immunity. Nature 548, 461–465. https://doi.org/10.1038/nature23449

Magis, A. De, 2016. G-quadruplex binders cause DNA damage by inducing R-loops in human cancer cells. alma.

Maizels, N., 2015. G4-associated human diseases. EMBO Rep. 16, 910–922. https://doi.org/10.15252/embr.201540607

Manzo, S.G., Hartono, S.R., Sanz, L.A., Marinello, J., De Biasi, S., Cossarizza, A., Capranico, G., Chedin, F., 2018. DNA Topoisomerase I differentially modulates R-loops across the human genome. Genome Biol. 19, 100. https://doi.org/10.1186/s13059-018-1478-1

Mao, S.Q., Ghanbarian, A.T., Spiegel, J., Martínez Cuesta, S., Beraldi, D., Di Antonio, M., Marsico, G., Hänsel-Hertsch, R., Tannahill, D., Balasubramanian, S., 2018. DNA G-quadruplex structures mold the DNA methylome. Nat. Struct. Mol. Biol. 25, 951–957. https://doi.org/10.1038/s41594-018-0131-8

Marsico, G., Chambers, V.S., Sahakyan, A.B., McCauley, P., Boutell, J.M., Antonio, M. Di, Balasubramanian, S., 2019. Whole genome experimental maps of DNA G-quadruplexes in multiple species. Nucleic Acids Res. 47, 3862–3874. https://doi.org/10.1093/nar/gkz179

McKinney, W., 2010. Data Structures for Statistical Computing in Python. Proc. 9th Python Sci. Conf. 1697900, 51–56.

Medzhitov, R., Janeway, C.A., 1999. Innate immune induction of the adaptive immune response. Cold Spring Harb. Symp. Quant. Biol. 64, 429–435. https://doi.org/10.1101/sqb.1999.64.429

Nadel, J., Athanasiadou, R., Lemetre, C., Wijetunga, N.A., Ó Broin, P., Sato, H., Zhang, Z., Jeddeloh, J., Montagna, C., Golden, A., Seoighe, C., Greally, J.M., 2015. RNA:DNA hybrids in the human genome have distinctive nucleotide characteristics, chromatin composition, and transcriptional relationships. Epigenetics and Chromatin 8.

https://doi.org/10.1186/s13072-015-0040-6

Okazaki, T., Honjo, T., 2007. PD-1 and PD-1 ligands: From discovery to clinical application. Int. Immunol. https://doi.org/10.1093/intimm/dxm057

Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.C., Mendell, J.T., Salzberg, S.L., 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat. Biotechnol. 33, 290–295. https://doi.org/10.1038/nbt.3122

Phoenix, P., Raymond, M.A., Massé, É., Drolet, M., 1997. Roles of DNA topoisomerases in the regulation of R-loop formation in vitro. J. Biol. Chem. 272, 1473–1479. https://doi.org/10.1074/jbc.272.3.1473

Quinlan, A.R., Hall, I.M., 2010. BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842. https://doi.org/10.1093/bioinformatics/btq033

Ramírez, F., Ryan, D.P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A.S., Heyne, S., Dündar, F., Manke, T., 2016. deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res. 44, W160–W165. https://doi.org/10.1093/nar/gkw257

Ratmeyer, L., Zhong, Y.Y., Wilson, W.D., Vinayak, R., Zon, G., 1994. Sequence Specific Thermodynamic and Structural Properties for DNA·RNA Duplexes. Biochemistry 33, 5298–5304. https://doi.org/10.1021/bi00183a037

Reaban, M.E., Lebowitz, J., Griffin, J.A., 1994. Transcription induces the formation of a stable RNA·DNA hybrid in the immunoglobulin α switch region. J. Biol. Chem. 269, 21850–21857.

Read, M., Harrison, R.J., Romagnoli, B., Tanious, F.A., Gowan, S.H., Reszka, A.P., Wilson, W.D., Kelland, L.R., Neidle, S., 2001. Structure-based design of selective and potent G quadruplex-mediated telomerase inhibitors. Proc. Natl. Acad. Sci. U. S. A. 98, 4844–4849. https://doi.org/10.1073/pnas.081560598

Repana, D., Nulsen, J., Dressler, L., Bortolomeazzi, M., Venkata, S.K., Tourna, A., Yakovleva, A., Palmieri, T., Ciccarelli, F.D., 2019. The Network of Cancer Genes (NCG): a comprehensive catalogue of known and candidate cancer genes from cancer sequencing screens. Genome Biol. 20, 1. https://doi.org/10.1186/s13059-018-1612-0

Richardson, J.P., 1975. Attachment of nascent RNA molecules to superhelical DNA. J. Mol. Biol. 98, 565–579. https://doi.org/10.1016/S0022-2836(75)80087-8

Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., Mesirov, J.P., 2011. Integrative genomics viewer. Nat. Biotechnol. https://doi.org/10.1038/nbt.1754

Rodriguez, R., Miller, K.M., Forment, J. V., Bradshaw, C.R., Nikan, M., Britton, S., Oelschlaegel, T., Xhemalce, B., Balasubramanian, S., Jackson, S.P., 2012. Small-molecule-induced DNA damage identifies alternative DNA structures in human genes. Nat. Chem. Biol. 8, 301–310. https://doi.org/10.1038/nchembio.780

Rodriguez, R., Müller, S., Yeoman, J.A., Trentesaux, C., Riou, J.F., Balasubramanian, S., 2008. A novel small molecule that alters shelterin integrity and triggers a DNA-damage response at telomeres. J. Am. Chem. Soc. 130, 15758–15759. https://doi.org/10.1021/ja805615w

Roy, D., Lieber, M.R., 2009. G Clustering Is Important for the Initiation of Transcription-Induced R-Loops In Vitro, whereas High G Density without Clustering Is Sufficient Thereafter. Mol. Cell. Biol. 29, 3124–3133. https://doi.org/10.1128/mcb.00139-09

Russo, M., De Lucca, B., Flati, T., Gioiosa, S., Chillemi, G., Capranico, G., 2019. DROPA: DRIP-seq optimized peak annotator. BMC Bioinformatics 20, 414. https://doi.org/10.1186/s12859-019-3009-9

Sahakyan, A.B., Chambers, V.S., Marsico, G., Santner, T., Di Antonio, M., Balasubramanian, S., 2017. Machine learning model for sequence-driven DNA G-quadruplex formation. Sci. Rep. 7. https://doi.org/10.1038/s41598-017-14017-4

Sanz, L.A., Hartono, S.R., Lim, Y.W., Steyaert, S., Rajpurkar, A., Ginno, P.A., Xu, X., Chédin, F., 2016. Prevalent, Dynamic, and Conserved R-Loop Structures Associate with Specific Epigenomic Signatures in Mammals. Mol. Cell 63, 167–178. https://doi.org/10.1016/j.molcel.2016.05.032

Schwalb, B., Michel, M., Zacher, B., Hauf, K.F., Demel, C., Tresch, A., Gagneur, J., Cramer, P., 2016. TT-seq maps the human transient transcriptome. Science (80-.). 352, 1225–1228. https://doi.org/10.1126/science.aad9841

Sergushichev, A.A., 2016. An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. bioRxiv 060012. https://doi.org/10.1101/060012

Shelby, M.D., 1988. The genetic toxicity of human carcinogens and its implications. Mutat. Res. Toxicol. 204, 3–15. https://doi.org/10.1016/0165-1218(88)90113-9

Shivji, M.K.K., Renaudin, X., Williams, Ç.H., Venkitaraman, A.R., 2018. BRCA2 Regulates Transcription Elongation by RNA Polymerase II to Prevent R-Loop Accumulation. Cell Rep. 22, 1031–1039. https://doi.org/10.1016/j.celrep.2017.12.086

Siddiqui-Jain, A., Grand, C.L., Bearss, D.J., Hurley, L.H., 2002. Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress c-MYC transcription. Proc. Natl. Acad. Sci. U. S. A. 99, 11593–11598. https://doi.org/10.1073/pnas.182256799

Skourti-Stathaki, K., Kamieniarz-Gdula, K., Proudfoot, N.J., 2014. R-loops induce repressive chromatin marks over mammalian gene terminators. Nature 516, 436–439. https://doi.org/10.1038/nature13787

Skourti-Stathaki, K., Proudfoot, N.J., Gromak, N., 2011. Human Senataxin Resolves RNA/DNA Hybrids Formed at Transcriptional Pause Sites to Promote Xrn2-Dependent Termination. Mol. Cell 42, 794–805. https://doi.org/10.1016/j.molcel.2011.04.026

Sollier, J., Stork, C.T., García-Rubio, M.L., Paulsen, R.D., Aguilera, A., Cimprich, K.A., 2014. Transcription-Coupled Nucleotide Excision Repair Factors Promote R-Loop-Induced Genome Instability. Mol. Cell 56, 777–785. https://doi.org/10.1016/j.molcel.2014.10.020

Soneson, C., Love, M.I., Robinson, M.D., 2015. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. F1000Research 4, 1521. https://doi.org/10.12688/f1000research.7563.1

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., Mesirov, J.P., 2005. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. U. S. A. 102, 15545–15550. https://doi.org/10.1073/pnas.0506580102

Sun, D., Hurley, L.H., 2009. The importance of negative superhelicity in inducing the formation of G-quadruplex and i-motif structures in the c-Myc promoter: Implications for drug targeting and control of gene expression. J. Med. Chem. 52, 2863–2874. https://doi.org/10.1021/jm900055s

Sun, H., Karow, J.K., Hickson, I.D., Maizels, N., 1998. The Bloom's syndrome helicase unwinds G4 DNA. J. Biol. Chem. 273, 27587–27592. https://doi.org/10.1074/jbc.273.42.27587

Sun, L., Wu, J., Du, F., Chen, X., Chen, Z.J., 2013. Cyclic GMP-AMP synthase is a cytosolic DNA sensor that activates the type I interferon pathway. Science (80-.). 339, 786–791. https://doi.org/10.1126/science.1232458

Thorsson, Vésteinn, Gibbs, D.L., Brown, S.D., Wolf, D., Bortone, D.S., Ou Yang, T.H., […], Westervelt, P., Rubin, M.A., Lee, J. Il, Aredes, N.D., Mariamidze, A., Serody, J.S., Demicco, E.G., Disis, M.L., Vincent, B.G., Shmulevich, Ilya, 2018. The Immune Landscape of Cancer. Immunity 48, 812-830.e14. https://doi.org/10.1016/j.immuni.2018.03.023

Wahba, L., Amon, J.D., Koshland, D., Vuica-Ross, M., 2011. RNase H and Multiple RNA Biogenesis Factors Cooperate to Prevent RNA:DNA Hybrids from Generating Genome Instability. Mol. Cell 44, 978–988. https://doi.org/10.1016/j.molcel.2011.10.017

Wahba, L., Costantino, L., Tan, F.J., Zimmer, A., Koshland, D., 2016. S1-DRIP-seq identifies high expression and polyA tracts as major contributors to R-loop formation. Genes Dev. 30, 1327–1338. https://doi.org/10.1101/gad.280834.116

Watson, J.D., Crick, F.H.C., 1953. Molecular Structures of Nucleic Acids. Nature 171, 737–738. https://doi.org/10.1038/171737a0

Wellinger, R.E., Prado, F., Aguilera, A., 2006. Replication Fork Progression Is Impaired by Transcription in Hyperrecombinant Yeast Cells Lacking a Functional THO Complex. Mol. Cell. Biol. 26, 3327–3334. https://doi.org/10.1128/mcb.26.8.3327-3334.2006

Westover, K.D., 2004. Structural Basis of Transcription: Separation of RNA from DNA by RNA Polymerase II. Science (80-.). 303, 1014–1016. https://doi.org/10.1126/science.1090839

Woo, S.R., Fuertes, M.B., Corrales, L., Spranger, S., Furdyna, M.J., Leung, M.Y.K., Duggan, R., Wang, Y., Barber, G.N., Fitzgerald, K.A., Alegre, M.L., Gajewski, T.F., 2014. STING-

dependent cytosolic DNA sensing mediates innate immune recognition of immunogenic tumors. Immunity 41, 830–842. https://doi.org/10.1016/j.immuni.2014.10.017

Xiao, S., Zhang, J.Y., Zheng, K.W., Hao, Y.H., Tan, Z., 2013. Bioinformatic analysis reveals an evolutional selection for DNA:RNA hybrid G-quadruplex structures as putative transcription regulatory elements in warm-blooded animals. Nucleic Acids Res. 41, 10379–10390. https://doi.org/10.1093/nar/gkt781

Xu, W., Xu, H., Li, K., Fan, Y., Liu, Y., Yang, X., Sun, Q., 2017. The R-loop is a common chromatin feature of the Arabidopsis genome. Nat. Plants 3, 704–714. https://doi.org/10.1038/s41477-017-0004-x

Yang, Y., McBride, K.M., Hensley, S., Lu, Y., Chedin, F., Bedford, M.T., 2014. Arginine Methylation Facilitates the Recruitment of TOP3B to Chromatin to Prevent R Loop Accumulation. Mol. Cell 53, 484–497. https://doi.org/10.1016/j.molcel.2014.01.011

Zhang, C.Z., Spektor, A., Cornils, H., Francis, J.M., Jackson, E.K., Liu, S., Meyerson, M., Pellman, D., 2015. Chromothripsis from DNA damage in micronuclei. Nature 522, 179–184. https://doi.org/10.1038/nature14493

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nussbaum, C., Myers, R.M., Brown, M., Li, W., Shirley, X.S., 2008. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 9, R137. https://doi.org/10.1186/gb-2008-9-9-r137

Zimmer, J., Tacconi, E.M.C., Folio, C., Badie, S., Porru, M., Klare, K., Tumiati, M., Markkanen, E., Halder, S., Ryan, A., Jackson, S.P., Ramadan, K., Kuznetsov, S.G., Biroccio, A., Sale, J.E., Tarsounas, M., 2016. Targeting BRCA1 and BRCA2 Deficiencies with G-Quadruplex-Interacting Compounds. Mol. Cell 61, 449–460. https://doi.org/10.1016/j.molcel.2015.12.004

ANNEX 1

DNA damage and genome instability by G-quadruplex ligands are mediated by R loops in human cancer cells

Alessio De Magis[a,1,2], Stefano G. Manzo[a,1,3], Marco Russo[a], Jessica Marinello[a], Rita Morigi[a], Olivier Sordet[b], and Giovanni Capranico[a,4]

[a]Department of Pharmacy and Biotechnology, Alma Mater Studiorum University of Bologna, 40126 Bologna, Italy; and [b]Cancer Research Center of Toulouse, INSERM, Université de Toulouse, Université Toulouse III Paul Sabatier, CNRS, 31037 Toulouse, France

Edited by Philip C. Hanawalt, Stanford University, Stanford, CA, and approved December 3, 2018 (received for review June 16, 2018)

G quadruplexes (G4s) and R loops are noncanonical DNA structures that can regulate basic nuclear processes and trigger DNA damage, genome instability, and cell killing. By different technical approaches, we here establish that specific G4 ligands stabilize G4s and simultaneously increase R-loop levels within minutes in human cancer cells. Genome-wide mapping of R loops showed that the studied G4 ligands likely cause the spreading of R loops to adjacent regions containing G4 structures, preferentially at 3′-end regions of expressed genes, which are partially ligand-specific. Overexpression of an exogenous human RNaseH1 rescued DNA damage induced by G4 ligands in *BRCA2*-proficient and *BRCA2*-silenced cancer cells. Moreover, even if the studied G4 ligands increased noncanonical DNA structures at similar levels in nuclear chromatin, their cellular effects were different in relation to cell-killing activity and stimulation of micronuclei, a hallmark of genome instability. Our findings therefore establish that G4 ligands can induce DNA damage by an R loop-dependent mechanism that can eventually lead to different cellular consequences depending on the chemical nature of the ligands.

R loop | G-quadruplex ligand | genome instability | DNA cleavage | antitumor activity

G quadruplexes (G4s) are noncanonical secondary DNA structures constituted of two or more stacked guanine tetrads held together by Hoogsteen hydrogen bonds and stabilized by monovalent cations such as K^+ and Na^+ (1, 2). G4s can play a regulatory role in basic nuclear functions such as replication and transcription, and indeed G4-promoting sequences have been mapped at key regulatory genomic sites, notably oncogene promoters, untranslated exonic regions, replication origins, and telomeres (1, 2). In the past years, several specific G4 ligands have been developed targeting telomeres or oncogene promoters, as G4s are considered promising targets of effective anticancer drugs (1, 2). Nevertheless, despite the high number of G4 ligands in the literature, few have entered early phases of clinical trials and none has shown efficacy in cancer patients (1–3).

An intriguing effect of G4 ligands is the induction of DNA damage and genome instability. In particular, pyridostatin (PDS) (*SI Appendix*, Fig. S1*A*), a well-known G4 ligand (1, 4), induces DNA damage as shown by formation of γH2AX foci (5), a marker of double-stranded DNA breakage (DSB). The compound triggers the activation of the DNA damage response (DDR) pathway, as determined by phosphorylation of ATM, DNA-PKcs, Chk1, and other factors and by cell-cycle arrest at G2/M phase (5). G4 ligands, including PDS, were recently shown to be more active in reducing the proliferation of *BRCA1/2*-deficient cancer cells by accumulating DNA damage, chromosomal aberrations, and persistent checkpoint activation (6, 7). These findings are consistent with a critical role of the homologous recombination repair (HRR) pathway in protecting cancer cells from genome instability triggered by G4 ligand activity. Consistently, G4 structures can lead to instability of the CEB1 minisatellite in *pif*Δ *Saccharomyces cerevisiae* cells in a manner dependent on HRR (8). G4 ligands can also induce genome instability showing specific gene interactions in different cell systems. For instance, the compound TMP_yP4, known to bind to telomere G4s, has been shown to enhance murine telomere fragility in the absence of RTEL1, a factor regulating the disassembly of telomeric T loops (a lasso-like telomere organization) (9). Recent work has shown that G4 structures can cause a high rate of sister chromatid exchange in Bloom helicase (*BLM*)-mutated cells derived from Bloom syndrome patients (10). The authors proposed that *BLM* preserves genome stability by resolving G4 structures and suppressing recombination at transcribed genomic loci. Thus, stabilization of G4s by specific ligands or genetic defects can lead to genome instability through the induction of DSB and/or activation of recombination repair pathways. Nevertheless, the mechanism of DSB formation and genome instability by G4 ligands is unknown.

A G4 can be structurally compatible with an R loop, which is another noncanonical secondary DNA structure wherein the two strands of a DNA duplex are separated and one of them is annealed to an RNA, forming a DNA:RNA hybrid (11–14). G4s were shown to form in the displaced strand of an R loop, forming a G loop, depending on high transcription rate and negative supercoiling of

Significance

In the past decades, several compounds have been developed specifically targeting G quadruplexes (G4s), as these noncanonical DNA structures are considered promising targets of effective anticancer drugs. However, despite the high number of known ligands, none showed efficacy in cancer patients. We have investigated the interplay of G4s with R loops, another noncanonical DNA structure, and the findings reveal a mechanism of genome instability and cell killing by G4 ligands, particularly effective in cancer cells deficient of *BRCA2* activity. This knowledge establishes a mechanistic model of G4 ligand activity in cancer cells that can open new lines of investigation aiming at developing clinically effective G4 ligands.

Author contributions: G.C. conceived the research; A.D.M., S.G.M., O.S., and G.C. designed research; A.D.M., S.G.M., M.R., J.M., and R.M. performed research; A.D.M., S.G.M., M.R., J.M., O.S., and G.C. analyzed data; and G.C. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: The data reported in this paper have been deposited in the Gene Expression Omnibus (GEO) database, https://www.ncbi.nlm.nih.gov/geo (accession no. GSE115957).

[1]A.D.M. and S.G.M. contributed equally to this work.

[2]Present address: Department of Oncology, Hematology and Rheumatology, University Hospital Bonn, 53127 Bonn, Germany.

[3]Present address: Division of Gene Regulation, Netherlands Cancer Institute, 1066 CX Amsterdam, The Netherlands.

[4]To whom correspondence should be addressed. Email: giovanni.capranico@unibo.it.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1810409116/-/DCSupplemental.

Published online December 27, 2018.

the DNA template (15). The structural compatibility of G4s and R loops is consistent with the knowledge that the formation of both G4s and R loops is favored by similar DNA structural aspects, such as G richness of displaced strands and negative torsional tension, which are common features of active gene promoters (16–18). Interestingly, R loops play a role in several physiological functions of cells; however, unscheduled R loops can lead to DSB, genome instability, and cell killing (12, 13, 19).

Thus, we have here investigated the effects of G4 ligands on R-loop formation and genome integrity in human cancer cells. By studying three structurally unrelated G4 ligands and an inactive analog, our findings establish that G4 ligands induce an immediate increase of nuclear R loops that mediate the formation of DSB. We also discovered that G4 ligands cause the generation of micronuclei at later times in an R loop-mediated manner, particularly in *BRCA2*-depleted cancer cells. Our findings establish a mechanistic role for R loops in mediating the cellular effects of G4 ligands, and open unexpected lines of investigation and development of new anticancer strategies.

Results

G4 Ligands Induce an Increase of Nuclear DNA:RNA Hybrid Structures. We set out to define the interactions of G4s with R-loop structures in relation to genome integrity in human U2OS cancer cells. We first determined with immunofluorescence microscopy (IF) the induction of DNA:RNA hybrids by three established and structurally different G4 ligands: pyridostatin (2), Braco-19 (2), and FG (compound **1** in refs. 20 and 21) (*SI Appendix*, Fig. S1*A*). Nuclear G4s and hybrids were visualized with BG4 and S9.6 antibodies, respectively, validated previously (21) or with specific assays (*SI Appendix*, Fig. S1 *B–G*). In particular, our high-stringency buffer conditions prevented the binding of S9.6 Ab to the cytoplasm, as we rarely detected cytoplasmic signals (Fig. 1; see also *SI Appendix*, Figs. S1, S3, and S4). G4 ligands robustly increased the number of nuclear G4s and hybrid foci between 2 and 10 min in U2OS cells, whereas they dropped close to baseline levels or lower at later times (30 to 60 min; Fig. 1 *A* and *B*). The kinetics of hybrid and G4 formation paralleled each other closely (Fig. 1*B*), and increased hybrids were located in the nucleoplasm, clearly outside the nucleolus, as visualized with nucleolin staining, indicating that they were not restricted to highly transcribed ribosomal RNA genes (Fig. 1*A*). The transient increase of G4s and hybrids at short times was specific, as an FG analog, FA (*SI Appendix*, Fig. S1*A*), which did not stabilize G4s in vitro and in living cells (compound **3** in ref. 20 and compound **14a** in ref. 21), did not increase hybrids either (Fig. 1 *A* and *B*). We must note that FA is more cytotoxic than FG (21) (see below), and thus the mechanism of action of the former is likely different from the latter.

As G4 focus stabilization by specific ligands was often reported to occur after 24 h (1, 2), we also tested these conditions and showed that PDS and FG, but not FA, also induced both G4 and

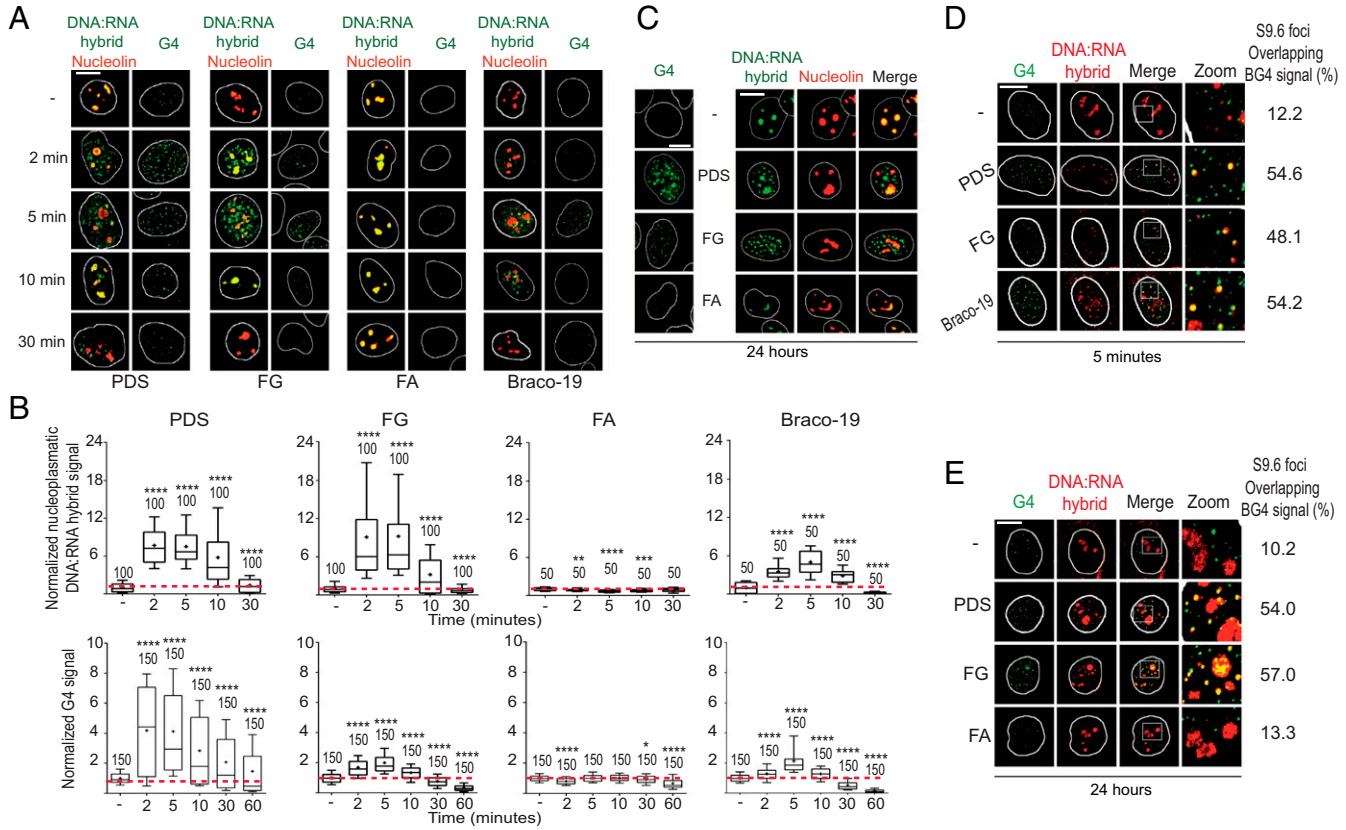

Fig. 1. G4 binders induce nuclear DNA:RNA hybrids overlapping G4 foci. (*A*) Human U2OS cells were treated with PDS, FG, FA (10 μM), or Braco-19 (15 μM) for the indicated times. IF images were analyzed after labeling G4s, hybrids, and nucleolin with BG4, S9.6, and AB22758 antibodies, respectively (as indicated by color). White lines indicate nuclei. (*B*) Hybrid and G4 levels were determined by fluorescence intensity (FI) of cells treated as in *A*. FI of the nucleoplasmic compartment was calculated by subtracting the nucleolar signal from total nuclear FI. The graphs show FI levels normalized over untreated cells of two biological replicates, and numbers indicate analyzed nuclei. Boxplots are as detailed in *SI Appendix, Methods*; horizontal lines and plus signs are median and mean values, respectively. Asterisks indicate statistical significance in comparison with untreated cells by the Kolmogorov–Smirnov parametric test. $*P < 0.05$, $**P < 0.01$, $***P < 0.001$, $****P < 0.0001$. (*C*) G4 foci and hybrid signals induced by 24-h treatments with 10 μM PDS, FG, or FA. U2OS cells were stained as in *A*. Hybrid and G4 levels are shown in *SI Appendix*, Fig. S2C. (*D*) Colocalization of hybrid signals with G4 foci. Cells were treated as in *A* for 5 min and then stained with BG4 (green) and S9.6 (red) antibodies. (*E*) Colocalization of hybrid signals with G4 foci as in *D*, but cells were treated for 24 h. (Scale bars, 10 μm.)

hybrid foci at 24 h (Fig. 1*C*). As we reported that the topoisomerase I (Top1) poison camptothecin can transiently enhance nuclear DNA:RNA hybrids at short times in cancer cells (22), we wondered whether FG and PDS can also poison Top1. We then measured Top1–DNA cleavage complexes in U2OS cells, as described (23). The results showed that PDS and FG are not Top1 poisons (*SI Appendix*, Fig. S2*A*), thus excluding the possibility that Top1 poisoning accounts for the increase of hybrids by the studied ligands. In addition, FG and FA were previously shown to have negligible binding activity toward DNA duplexes (20, 21). Thus, the cellular effects of the studied G4 ligands on hybrid induction are likely due to their specific G4 binding activity. In addition, while FG was as effective as PDS in increasing hybrid foci in the nucleoplasm (Fig. 1 *B* and *C*), PDS increased hybrid signals at nucleoli more than FG and Braco-19 (*SI Appendix*, Fig. S2 *B* and *C*). These observations therefore suggest that PDS, Braco-19, and FG may differently affect R loops along the genome of U2OS cells, likely due to binding to different sets of G4 targets.

Further investigations of the kinetics and dose dependence of hybrid and G4 formation by PDS, Braco-19, and FG showed that the increase of hybrids observed from 2 to 10 min was followed by several hours (0.5 to 6) with no induction, and then by a second increase at 18 to 24 h (*SI Appendix*, Fig. S3 *A*–*E*). In addition, hybrid induction was clearly dose-dependent for PDS and Braco-19 (*SI Appendix*, Fig. S3 *F* and *G*). We noted, however, that Braco-19 had somewhat different kinetics at 18 to 24 h in comparison with PDS and FG (*SI Appendix*, Fig. S3 *B* and *E*), suggesting different cellular outcomes among the compounds.

As the induction of hybrids was always coupled to increased G4 foci, our IF data are consistent with a direct effect of ligand-stabilized G4s on R-loop formation and/or stability. Moreover, the observed biphasic kinetics supports that nuclear R loops are highly dynamic structures (24), likely regulated by homeostatic mechanisms. In this context, G4 stabilization may act as a favoring factor that would, however, stimulate a counterbalancing factor that will then reduce R-loop levels. For instance, unscheduled R loops are expected to inhibit transcription, which would then disfavor the formation of G4s and R loops after the initial increase. Thus, we wondered whether G4s and R loops were localized in the same chromatin domain, and then performed colabeling confocal IF experiments with BG4 and S9.6 antibodies. Interestingly, hybrid signals significantly overlapped with G4 foci in cells treated for short (5 min) or long times (24 h) with PDS, Braco-19, and FG (Fig. 1 *D* and *E*). It must also be noted that several G4 foci did not overlap with hybrids (Fig. 1 *D* and *E*). Within the resolution limits of IF, these observations showed that G4 ligands may stimulate both hybrid and G4 foci at the same or very close chromatin domains and that stabilized G4s may favor a nearby R loop.

Next, we wondered whether G4 ligands can increase R-loop levels in other human cell lines. Since PDS has been shown to be less effective in G4 stabilization in normal cells (1, 25), we also determined G4 ligand effects either in normal human WI-38 and IMR-90 lung fibroblasts or HeLa cancer cells. Interestingly, PDS and FG did not increase G4s nor hybrids at detectable levels in normal WI-38 and IMR-90 fibroblasts at all time points, whereas the ligands increased G4 foci and hybrid signals in HeLa cells after 5-min and 24-h treatments (*SI Appendix*, Fig. S4). Thus, the data may suggest that the studied cancer cells may suffer from a loss of function(s) resulting in a defect in the removal of G4 structures and hence in detectable IF signals. Altogether, our results show that the studied G4 ligands can induce the simultaneous formation of G4 and DNA:RNA hybrid structures at close chromatin domains in the studied human cancer cells.

G4 Ligands Induce R-Loop Spreading into Adjacent Regions Containing Experimentally Observed G4 Structures. To gain insights into the mechanism of R-loop induction by G4 ligands, we wondered whether genomic locations of R loops overlapped with G4 structures, as previously established in human genomic DNA in the presence of PDS with a polymerase-stop assay (26). Thus, we focused on two G4 ligands, PDS and FG, and determined genomic R-loop maps by DRIP-seq (DNA:RNA immunoprecipitation-using sequencing) (18, 27) in U2OS cells treated for 5 min with the compounds to identify the genomic sites of affected R loops. Two biological replicates were performed for untreated and treated cells. To identify specifically the hybrids, we sequenced recovered DNAs from those cell samples that had been left untreated or treated with *Escherichia coli* RNaseH after restriction enzyme digestion and before immunoprecipitation with S9.6 (Fig. 2*A*; see also *SI Appendix*, *Methods*). R-loop peaks were then identified only if they were consistently observed in both replicates and absent in the RNaseH-treated samples. Fig. 2*A* shows a representative gene, TLE3 (Transducin-Like Enhancer of Split 3), which encodes a transcriptional corepressor protein. With these stringent criteria, we obtained thousands of R-loop peaks in control and treated cells covering from 2.5 to 5.1% of the genome (*SI Appendix*, Fig. S5*A*), and each pair of biological replicates showed high correlation coefficients (*SI Appendix*, Fig. S5*B*). R-loop peaks were consistently found in gene regions and were highly enriched at 5′- and 3′-end gene regions (*SI Appendix*, Fig. S5*C*), in agreement with previous findings (18, 24, 27–30). We observed that the profiles of R-loop peaks were highly correlated with each other, and genomic peak distributions were very similar between control and treated cells (*SI Appendix*, Fig. S5 *D* and *E*). However, the peak number and genome coverage were higher for treated than control cells (*SI Appendix*, Fig. S5*A*), suggesting an increase of R loops by the two ligands without alterations of global patterns of genomic R loops.

As the observed genomic increase can be due to higher R-loop levels at specific regions or to the spreading of preexisting peaks, we then investigated both possibilities. A direct comparison of peak intensity showed a high number (97%) of increased peaks (gain), whereas decreased peaks (loss) were only few (FG: 4,411 gain and 149 loss; PDS: 9,881 gain and 272 loss). Gain peaks were particularly enriched at the 3′ end of genes (Fig. 2*B*), but we did not observe a significant enrichment of experimentally observed G4 motifs (26) with gain peaks compared with unchanged peaks. However, as R loops and G4s have been associated with active transcription and promoters (18, 27, 28, 31), we then measured gene expression levels in control cells by RNA-seq to determine transcription-dependent effects of G4 ligands on R-loop levels. Then, we divided genes into four classes depending on transcription levels, and calculated the increase of DRIP sequence reads induced (as Δ reads) by PDS and FG at their transcription start sites for each expression category (Fig. 2*C*). The results show that the increase of DRIP reads is highly correlated with transcription levels and the presence of CpG islands, thus suggesting that both transcription and guanine-rich sequences can favor the increase of R loops, likely due to a prompt binding of the ligands to their targets at active and accessible promoters. Moreover, we wondered whether GC skew (G richness on the nontranscribed strand) could affect R-loop increase by the studied G4 ligands. The data show that the studied ligands increased DRIP reads at higher levels in actively transcribed CpG-island promoters with GC skew than in those without GC skew (*SI Appendix*, Fig. S5*F*), further supporting a critical role for ligand binding to G4 targets at active promoters.

Next, as we noticed that R-loop peaks were often shared between control and G4 ligand-treated cells (see instances in Fig. 2*A*), we wondered whether gains were due to extended R loops more than to higher peak intensity. Thus, we analyzed the length of common peaks (more than 13,000 for each compound) and found that a significant number of them were extended by PDS and FG (Fig. 3*A*). Interestingly, extended peaks were enriched particularly at gene 3′ ends for both G4 ligands (Fig. 3*B*). This indicated that G4 ligands could frequently induce R-loop

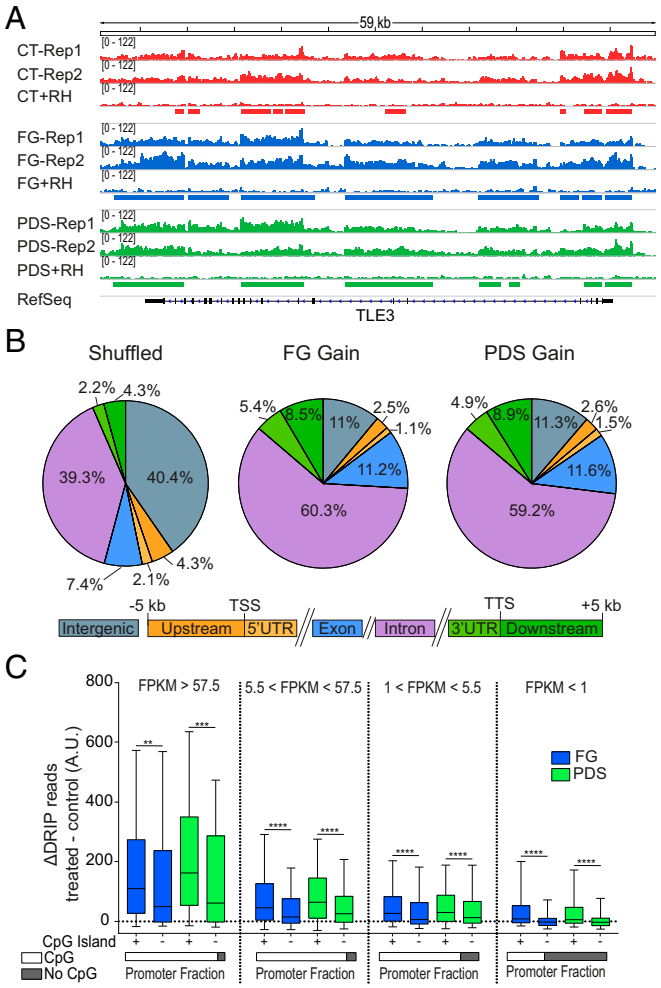

Fig. 2. G4 binders increase R-loop levels at CpG-island promoters of transcribed genes. (*A*) Genomic R-loop profiles at the TLE3 gene locus. Two biological replicates are shown for either control or G4 binder-treated cells along with an RNaseH-treated sample. Control cells are red; cells were treated with FG (blue) or PDS (green) for 5 min. Rectangles indicate R-loop peaks. (*B*) Distribution of DRIP peak gains for FG (*Middle*) and PDS (*Right*) across the genomic compartments depicted below. Numbers indicate the percentage occupied by each compartment. (*B*, *Left*) The graph shows the compartment distribution of randomly shuffled peaks over the full genome. TSS, transcription start site; TTS, transcription termination site. (*C*) Increased DRIP read counts at active promoter TSSs are dependent on the transcription level and the presence of CpG islands. Analyzed regions are from 2,000 bp upstream to 2,000 bp downstream of the TSS. Genes are split into four categories based on transcript levels as established by RNA-seq data and indicated at the top of the graphs. FPKM, fragments per kilobase of gene exon model per million reads mapped. As indicated below, TSS promoters with CG islands constitute 92.9, 90.6, 81.8, and 37.4% for the four gene sets from left to right, respectively. Statistical significance was determined with the Kolmogorov–Smirnov test. $^{**}P < 0.01$, $^{***}P < 0.001$, $^{****}P < 0.0001$.

spreading to adjacent regions. Thus, to understand whether R-loop spreading was associated with nearby G4s, we determined the overlapping between extended peak regions and experimentally observed G4 structures (26), focusing on extended peaks with a fold change >1.5 and $P < 0.05$ (1,000 and 619 for FG and PDS, respectively; red asterisks, Fig. 3*A*). To take into consideration the strand forming the G4 or the hybrid, we considered that the observed G4 dataset is constituted by over 700,000 G4 sequences assigned to one of the two genomic strands (26). Then, we assigned DNA:RNA hybrids mapped at transcribed genes to the template strand of genes (*SI Appendix*, *Methods*). As R-loop peaks are frequent in GC-rich sequences (Fig. 2*C*), we also selected peaks not extended by G4 ligands (unchanged; blue dots, Fig. 3*A*) and matched to extended peaks for length and gene localization to be compared with extended peaks. Then, we calculated the enrichment of observed G4s at extended and unchanged peaks relative to random peaks (*SI Appendix*, *Methods*). The results showed that G4s in the displaced strand of R loops were more enriched in extended peaks than unchanged peaks (darker vs. lighter colors, Fig. 3*C*) for both PDS and FG, whereas G4s in the template strand were not enriched (Fig. 3*C*). Interestingly, FG-extended peaks show a significant depletion of extensions without any G4 (Fig. 3*C*). Thus, the statistical analyses suggest that G4 ligands can induce R-loop spreading when G4 structures are present in the displaced strand of adjacent regions. A comparison of extended peaks by FG and PDS showed that 248 peaks only were in common between the two ligands, while a large fraction of them were ligand-specific (57 to 74%; Fig. 3*D*), supporting a degree of ligand binding specificity to distinct genomic sets of G4 targets. To validate the bioinformatic analyses and R-loop spreading, we performed DRIP-qPCR determinations of R-loop levels at 15 extended peaks in cells treated with PDS and FG for 5 min. All of the tested regions with one exception (*VSIG8* gene for PDS) showed an increase of R-loop levels by the two ligands (Fig. 3*E*). Thus, altogether, these findings provide evidence that a G4 structure in the displaced strand of an R loop can likely stabilize and extend the overall structure when bound by specific ligands in cancer cells.

G4 Ligand-Induced DNA Damage Is Mediated by R Loops. As the studied G4 ligands can stabilize G4s along with R loops in nuclear chromatin of human cancer cells, we next investigated the biological consequences of R-loop induction. In particular, as G4 ligands are known to induce DNA damage and genome instability (1, 2), we asked whether this is due to increased levels of R loops.

First, we assessed the induction of DNA damage by PDS and FG in U2OS cells. Following 24 h of treatment, the two ligands induced an increase of S139-phosphorylated histone H2AX (γH2AX) foci (Fig. 4*A*) and of G2/M cells (*SI Appendix*, Fig. S6*A*), which are both hallmarks of genomic DSB and DDR. Moreover, we detected a marked increase of foci of 53BP1 (p53-binding protein 1) and S1778-phosphorylated 53BP1 (p53BP1; a specific marker of DSB and DDR activation) in cells treated with PDS for 24 h (Fig. 4 *B* and *C* and *SI Appendix*, Fig. S6*B*). Interestingly, p53BP1 foci showed a nearly perfect colocalization with γH2AX foci (Fig. 4*B*). DNA damage checkpoint activation was also assessed by measuring the induction of pATM (S1981-phosphorylated ATM; a marker of DDR activation) by PDS after 24-h treatments (Fig. 4 *B–D* and *SI Appendix*, Fig. S6*C*). Interestingly, pATM foci fully colocalized with γH2AX foci (Fig. 4*B*), indicating ATM recruitment to chromatin sites of DSB. FG and FA have minor effects on the levels of p53BP1 and pATM foci at 24 h (Fig. 4*D*, *Left* and *SI Appendix*, Fig. S6 *B–E*). However, the ratio pATM/ATM was increased by PDS and FG, but not FA, after 24 h (Fig. 4*D*, *Right*), suggesting that DDR is activated after 24 h with PDS as well as FG.

Then, as G4 ligands increased hybrid levels after 2 to 10 min (see above), we measured γH2AX focus levels at shorter times. γH2AX foci were consistently increased around twofold by PDS and FG after 1 to 4 h of treatment (Fig. 4*E*), in agreement with a published report on PDS (5), showing that the increase of unscheduled R loops preceded γH2AX focus formation. Interestingly, the γH2AX kinetics of PDS was markedly different from that of FG (Fig. 4*E*). In response to PDS, γH2AX focus number increased progressively over 24 h whereas, in response to FG, γH2AX foci reached a plateau after 2 to 4 h and then decreased somewhat after 20 to 24 h (Fig. 4*E*). FA, which did not induce G4s and R loops (see above), was slightly effective at inducing γH2AX, but less than FG (Fig. 4 *A*, *D*, and *E*). Braco-19

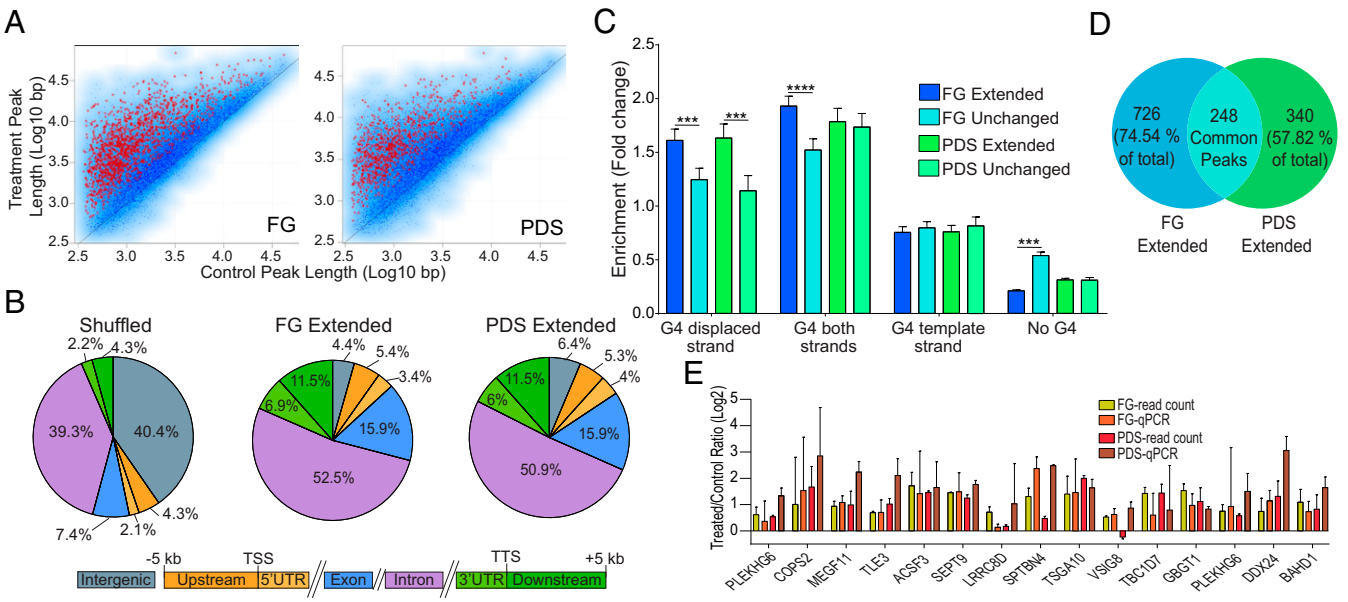

Fig. 3. G4 binders extend R loops at adjacent regions enriched for G4 motifs in the displaced strand. (*A*) Scatter plots of R-loop lengths of common peaks (13,539 and 13,316 for FG and PDS, respectively) between untreated and G4 binder-treated cells. Extended peaks with a length fold change >1.5 and $P < 0.05$ (red asterisks) are 1,000 and 619 for FG and PDS, respectively. Peaks with a length fold change <0.66 and $P < 0.05$ (not highlighted) are 8 and 14 for FG and PDS, respectively. Tests used were the *t* test and robust moderated *t* test from the limma R package (*SI Appendix, Methods*). (*B*) Distribution of DRIP extended peaks for FG (*Middle*) and PDS (*Right*) across the genomic compartments depicted below. Numbers indicate the percentage occupied by each compartment. (*B*, *Left*) The graph shows the compartment distribution of randomly shuffled peaks over the full genome. (*C*) Enrichments of experimentally established G4 motifs (29) in extended regions of extended peaks (red asterisks in *A*) in comparison with unchanged peaks (blue dots in *A*), matched to extended peaks for length and genomic localization. Extended peaks in genic regions only were considered for the analysis (974 for FG, 588 for PDS). Test used: Kolmogorov–Smirnov test. ***$P < 0.001$, ****$P < 0.0001$. (*D*) Venn diagram showing peak overlap between genic extended peaks for FG and PDS. (*E*) R-loop levels measured by DRIP-qPCR and DRIP-seq read count at 15 extended peaks. DRIP-qPCR measurements and DRIP-seq read count measurements in the same amplicon regions are shown as fold changes (treated over control). Bars show means ± SEM of four determinations from two experiments. Recovery of R loop-negative regions is around 100-fold less than positive loci (*SI Appendix*, Fig. S1).

also increased γH2AX foci, with a kinetics similar to that of PDS (*SI Appendix*, Fig. S6*F*). As G4 ligands did not stabilize G4s and hybrids in normal WI-38 fibroblasts (*SI Appendix*, Fig. S4), we also determined γH2AX levels in these cells. Consistently, G4 ligands did not induce γH2AX in WI-38 fibroblasts (*SI Appendix*, Fig. S6*G*).

Next, we asked whether DSB induced by G4 ligands is mediated by unscheduled R loops. To this end, we used a U2OS cell line which has stably been transfected with a vector expressing an mCherry-RNaseH1 under the control of a doxycycline (Doxy)-inducible Tet promoter (*SI Appendix*, Fig. S7*A*) (32). PDS induced a 1.8- to 2.1-fold increase of γH2AX foci in cells not expressing the enzyme, whereas PDS induced a 0.83-fold change in cells overexpressing mCherry-RNaseH1 (Fig. 4 *F* and *G*). The results demonstrate that RNaseH1 overexpression fully prevented the induction of γH2AX foci by PDS. Interestingly, PDS and FG induced low levels of G4 foci in cells overexpressing mCherry-RNaseH1 at short times, suggesting that RNaseH1 overexpression can prevent a full stabilization of G4 structures by the studied ligands (*SI Appendix*, Fig. S7*B*). Altogether, the results thus support that G4 ligand-induced DNA damage is mediated by unscheduled G4/R-loop structures.

To understand whether R-loop induction has any consequence on cell death induced by the studied G4 ligands, we determined the cytotoxic activity of PDS, FG, and FA in U2OS and U2OS_RH cells, the latter being a cell line stably transfected with a FLAG-tagged human RNaseH1 gene under a doxycycline-inducible Tet promoter (*SI Appendix*, Fig. S8*A*). Cell-killing activity of FG was reduced in U2OS_RH cells compared with U2OS cells, and the reduction was stronger when RNaseH1 was overexpressed by doxycycline, whereas cell-killing activity of FA was essentially unaffected (Table 1). PDS data were not meaningful, as it was poorly cytotoxic (Table 1). As FA did not increase G4s and R loops (Fig. 1) and is even more cytotoxic than FG, its mechanism of action is independent of the studied noncanonical DNA structures. Thus, overall, the findings support a main role for R loops in the induction of DNA damage and cell killing by the studied G4 ligands in human cancer cells.

G4 Ligand-Induced DNA Damage Is Mediated by R Loops in *BRCA2*-Depleted Cancer Cells. As G4 ligand-induced DSB can be repaired by HRR mechanisms and *BRCA2*-deficient cells are more sensitive to G4 ligands (6, 7), we wondered whether the hypersensitivity of HRR-deficient cells to G4 ligands was dependent on R loops. Therefore, we first determined whether the HRR pathway is activated in U2OS cancer cells by assessing foci formation of Rad51, a factor involved in the essential strand-invasion step of the HRR pathway (33, 34). IF results showed a consistent increase of Rad51 foci by 24-h treatments with PDS and FG, but not FA, to a very similar extent (Fig. 4*H* and *SI Appendix*, Fig. S8*B*), indicating an activated HRR in U2OS cells. Then, to establish the role of R loops in HRR-deficient cells, we silenced the *BRCA2* gene with siRNA in both U2OS and U2OS_RH cell lines (Fig. 5*A* and *SI Appendix*, Fig. S8*C*) and determined the number of γH2AX foci induced by PDS and FG with or without RNaseH1 overexpression.

Surprisingly, the effects of BRCA2 silencing were somewhat different between FG and PDS. In *BRCA2*-silenced U2OS cells, γH2AX focus levels by PDS were increased at early times compared with WT cells (from 1.26 to 1.54 fold change at 1 h, and from 1.98 to 2.69 fold change at 4 h) (Fig. 5*B* and *SI Appendix*, Fig. S9*A*). In contrast, the kinetics of γH2AX by FG was not altered and the levels of γH2AX foci were even somewhat reduced by *BRCA2* silencing (Fig. 5*C* and *SI Appendix*, Fig. S9*B*). Next, as γH2AX foci were increased at early times, we determined the

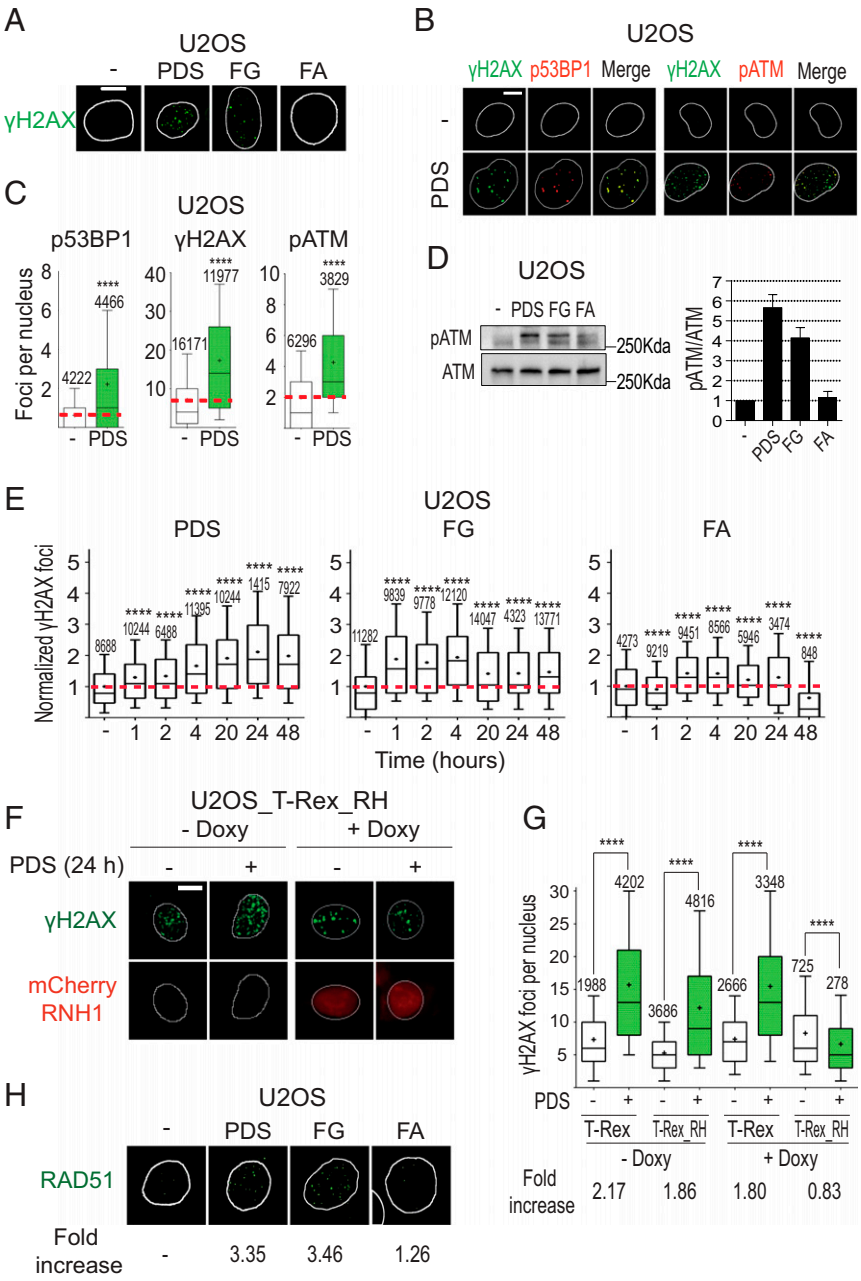

Fig. 4. PDS and FG induce DNA damage in an R loop-dependent manner. (*A*) γH2AX (S139-phosphorylated H2AX) foci induced by 10 μM PDS, FG, and FA following 24-h treatments. (*B*) Colocalization of γH2AX foci with p53BP1 (S1778-phosphorylated 53BP1) or pATM (S1981-phosphorylated ATM) foci in cells treated with 10 μM PDS for 24 h. Cells were costained for γH2AX (green) and p53BP1 (red) (*Left*), and for γH2AX (green) and pATM (red) (*Right*). (*C*) Levels of γH2AX, p53BP1, and pATM signals in cells treated as in *B*. (*D*) Western blots of pATM and ATM after 24-h treatments with 10 μM compounds (full membrane is shown in *SI Appendix*, Fig. S6*C*). (*D*, *Right*) Levels of the pATM/ATM ratio after treatment with the indicated compound. The graph shows mean values with standard errors of three biological replicates. (*E*) γH2AX focus levels in U2OS cells treated with the indicated compounds and for the indicated times. (*F*) PDS-induced γH2AX foci in cells expressing an exogenous RNaseH1. T-REX (control vector) and T-REX_RH (RNaseH1-expressing vector) stably transfected U2OS cells were treated with 10 μM PDS for 24 h with or without doxycycline, which activates RNaseH1 expression. RNaseH1 was fused to an mCherry tag to visualize cells with expressed enzyme. (*G*) γH2AX levels in T-REX and T-REX_RH cells treated (green bars) or not (white bars) with PDS as in *F*. Fold increase shows ratios of γH2AX levels in PDS-treated cells over corresponding untreated cells. (*H*) RAD51 foci induced by the indicated compounds (10 μM) after 24-h treatments of U2OS cells. (Scale bars, 10 μm.) All graphs show data from at least two biological replicates, reported as detailed in the legend to Fig. 1*B*. Asterisks indicate statistical significance in comparison with untreated cells by the Kolmogorov–Smirnov parametric test. ****$P < 0.0001$. (Magnification: *H*, 63×.)

induction of γH2AX foci following a 4-h treatment with the studied G4 ligands in U2OS_RH cells upon RNaseH1 expression by doxycycline. Similar results were observed in *BRCA2*-silenced and WT U2OS_RH cells in the absence of doxycycline (Fig. 5*D*, from 1.36 to 1.53 fold change; Fig. 5*E*, from 3.44 to 1.34 fold change), further supporting a difference in γH2AX induction between the two G4 ligands. However, exogenous RNaseH1 expression abolished the induction of γH2AX foci by either G4 ligands in *BRCA2*-silenced or WT cells (Fig. 5 *D* and *E*, + Doxy), showing a complete rescue of DSB. Thus, the findings strongly support that R loops play a main role in DSB induction by PDS and FG also in *BRCA2*-silenced cells regardless of any

Table 1. Exogenous RNaseH1 overexpression reduces cell-killing activity of FG but not FA

Compound	IC_{50}, μM*		
	U2OS	U2OS_RH	U2OS_RH + Doxy
FG	15.9 ± 1.2	52.8 ± 1.1	92.8 ± 1.1
FA	6.77 ± 1.4	7.10 ± 1.3	7.12 ± 1.3
PDS	>50	>50	>50

*Compound concentration inhibiting 50% of cell growth (see *SI Appendix, Methods*). Cells were exposed to the indicated compound for 24 h, and cell survival was determined after a further 48 h in drug-free medium. Numbers are means ± SD of two biological replicates, each performed in triplicate.

different molecular activity of the two ligands. In addition, the data suggest that the reported hypersensitivity of HRR-deficient cells to G4 ligands (6, 7) may be due to unscheduled R-loop formation.

PDS Induces Micronuclei Mediated by R-Loop Formation. During the course of this work, we observed that the studied G4 ligands could increase micronuclei, a clear hallmark of genome instability (35). As genome instability has been linked to impaired regulation of G4 structures in living cells (1, 8, 10, 36), we then asked whether micronucleus induction was mediated by DNA damage and R-loop/G4 structures. First, we investigated micronucleus induction in U2OS cells, showing that PDS increased the fraction of cells with micronuclei to a greater extent in *BRCA2*-silenced than *BRCA2* WT cells (Fig. 5*H* and *SI Appendix*, Fig. S9*C*), suggesting that the observed increase of DSB at early times (Fig. 5*B*) may lead to enhanced formation of micronucleated cells at later times. PDS-induced micronuclei were of different size (*SI Appendix*, Fig. S9 *C* and *D*), found in cytoplasmic regions close to the nucleus, and often showing IF signals of γH2AX (Fig. 5*G*), similar to recent reports using ionizing radiation (37–39). To better characterize micronucleus generation, we performed a cotreatment of PDS with a DNA-PK inhibitor, which fully abolished micronucleated *BRCA2*-silenced and WT U2OS cells (*SI Appendix*, Fig. S9 *E* and *F*) while maintaining a high level of γH2AX foci (*SI Appendix*, Fig. S9*G*), consistent with a strong inhibition of DNA repair. In agreement with previous reports (37–39), as the DNA-PK inhibitor can potentiate a cell-cycle G2/M arrest due to DNA-repair inhibition, the results showed that PDS can trigger micronuclei when PDS-induced DSB fails to be properly repaired and cells transit through mitosis.

We then determined whether micronucleus generation was related to unscheduled R loops. Interestingly, RNaseH1 overexpression in U2OS_RH cells abolished PDS induction of micronuclei in WT cells (from 1.57- to 0.74-fold; Fig. 5*I*) while reducing it in *BRCA2*-silenced cells (from 2.23- to 1.59-fold; Fig. 5*I*). Therefore, RNaseH1 could fully rescue micronucleus formation by PDS in WT cells but only partially in *BRCA2*-silenced cells. However, we noted that *BRCA2* silencing itself increased micronucleated cells (for instance, from 10.4 to 22.4% in untreated U2OS_RH cells without Doxy; Fig. 5*I*), and RNaseH1 overexpression somewhat affected micronucleus numbers as well (Fig. 5*I*). These observations may suggest that DNA repair or mitotic mechanisms of micronucleus generation may involve DNA:RNA hybrid formation, the removal of which might have an opposite effect on PDS-triggered micronuclei.

In contrast to PDS, FG increased the number of cells with micronuclei only slightly (*SI Appendix*, Fig. S10), suggesting that FG-induced DSB largely leads to a different molecular outcome such as cell-killing activity (Table 1). In addition, the slight induction of micronuclei did not allow establishing a clear effect by RNaseH1 overexpression on FG-induced micronuclei (*SI Appendix*, Fig. S10). Thus, the results show that PDS-induced DNA breaks are particularly prone to trigger micronuclei in a manner dependent on unscheduled R loops and G4 structures in *BRCA2* WT and, partially, in *BRCA2*-depleted cancer cells.

Discussion

R loops form abundantly in mammalian genomes and have been associated with different outcomes such as chromatin patterning, Ig gene recombination, DNA DSB, and genome instability (12, 13, 18, 24, 40, 41). However, how R-loop metabolism is regulated is still largely unknown. The present work provides experimental evidence that G4 structures can modulate the formation of R loops at active genes in eukaryotes. We here demonstrate that the studied G4 ligands induce unscheduled R-loop/G4 structures in human U2OS cancer cells likely by extending cotranscriptional R loops, which mediate the cellular activity of the studied compounds. Interestingly, even though PDS and FG can both increase R loops with similar kinetics, the biological outcome is partially different, in terms of DSB kinetics, genome instability, and cell-killing activity. In addition, we discovered that PDS can promote micronucleus generation in cancer cells in a manner dependent on unscheduled R loops. Our findings establish a molecular mechanism of G4 ligands with the potential to open new perspectives for the discovery and development of effective anticancer ligands.

Immediately upon cell exposure, G4 ligands increase R-loop levels and G4 foci, with a maximum at 2 to 10 min, after which R loops decline to undetectable levels for several hours. This immediate occurrence is likely due to the specific action of the compounds, namely stabilization of G4 structures through direct binding (Fig. 6). Thus, the simultaneous and rapid biphasic kinetics are a specific outcome of the studied G4 ligands in U2OS cells, likely due to a dynamic balance under homeostatic control of G4/R-loop levels. The immediate and rapid kinetics of R-loop/G4 structures may be due to a topological effect of G4 stabilization, as noncanonical DNA structures and DNA-duplex torsional stress may affect each other (16). On the other hand, the subsequent rapid reduction could result from transcription inhibition caused by increased R-loop levels and extensions (Figs. 2 and 3), and/or by active R-loop/G4 structure removal by repair mechanisms or specific helicases (1, 12, 14, 28). However, the mechanistic nature of the observed dynamic balance needs to be established in future work. Genomic maps showed that G4 ligands mainly induced R-loop gains at highly expressed genes, in particular at 5′- and 3′-end gene regions, after 5 min of treatment. The results are consistent with the findings that G4 structures are more often found at open chromatin sites of active genes in untreated cells (42). The reported genomic analyses suggest that PDS and FG can likely extend preexisting R loops to adjacent regions that are enriched for G4-promoting sequences in the displaced, but not template, strand of R loops. As our data are not at single-molecule levels but derive from statistical analyses of several genomic regions, we cannot exclude that increased (extended) R loops may be instead distinct R-loop structures. However, a likely hypothesis is that G4 ligands affect preexisting R loops within minutes of cell treatment, mainly at transcribed regions that are in an open chromatin conformation, in agreement with G4 structures present at these regions in untreated cells (42). Thus, G4 ligands may stabilize preexisting G4 structures in the displaced strand of R loops, in turn stabilizing DNA:RNA hybrids and increasing its length as proposed in our model (Fig. 6), in agreement with the G-loop model shown in *E. coli* (15). However, we believe that distinct mechanisms may be operative at functionally different chromatin regions, leading to reciprocal stabilization of R-loop/G4 structures. Thus, future studies will establish the precise mechanism at specific chromatin regions in living cells.

The increase of unscheduled R-loop/G4 structures can occur at different sets of transcribed genes depending on the specific ligand (Fig. 3*D*). Therefore, although the mechanistic model of

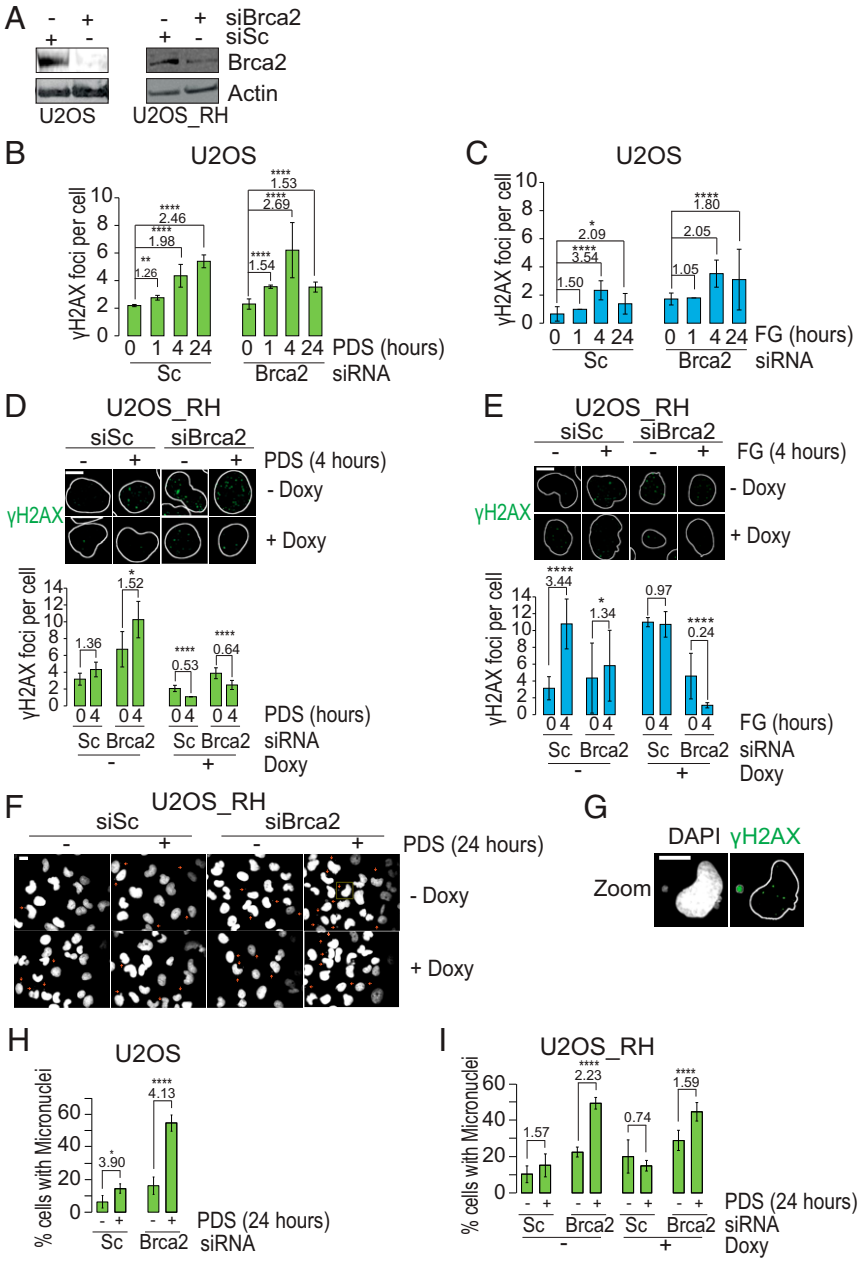

Fig. 5. DNA damage and genome instability are mediated by R loops in *BRCA2*-depleted cancer cells. (*A*) BRCA2 levels in the indicated cells following treatment with the specific siRNA against *BRCA2* (siBRCA2) or scrambled siRNA (siSc) for 48 h (full membranes are shown in *SI Appendix*, Fig. S8*C*). (*B*) γH2AX levels in *BRCA2*-silenced or WT U2OS cells treated with PDS for the indicated times. (*C*) Same as in *B* but cells were treated with FG. (*D*) PDS-induced γH2AX foci in *BRCA2*-silenced cells expressing an exogenous RNaseH1. RNaseH1-expressing vector stably transfected U2OS_RH cells were treated with siBRCA2 or siSc, and then with 10 μM PDS for 24 h with or without doxycycline to activate RNaseH1 expression. RNaseH1 was fused to a FLAG tag to detect the enzyme (*SI Appendix*, Fig. S8*A*). (*E*) Same as in *D* but cells were treated with FG. (*F*) Micronuclei induced in U2OS_RH cells transfected with scrambled siRNA or siRNA against *BRCA2*, and then treated with PDS for 24 h. Small red arrows indicate single micronuclei. (*G*) Enlargement of a cell in the yellow square in *F*. The image shows a micronucleus positive for γH2AX labeling. (*H*) Fractions of U2OS cells with micronuclei with and without *BRCA2* silencing and 24-h treatments with PDS. (*I*) Fractions of U2OS_RH cells with micronuclei with and without *BRCA2* silencing, doxycycline, and PDS treatments as indicated. (Scale bars, 10 μm.) Bars show mean values ± SEM. Fold-increase values are reported above the bars and represent treated/control ratios. Data in all panels are from at least two biological replicates, and in each experiment an average of 250 cells per sample was determined. Statistical significance was determined with the Kolmogorov–Smirnov test performed on the full cell populations. $*P < 0.05$, $**P < 0.01$, $****P < 0.0001$.

R-loop increase by PDS and FG can be similar, the cellular and molecular outcomes can be different depending on the genomic and chromatin context of the increased R loops (Fig. 6). We indeed report a number of differences between PDS and FG, mainly for γH2AX focus kinetics, BRCA2 silencing effects, micronucleus formation, and cell killing. As we used equimolar doses of the studied compounds, resulting in similar cellular levels of R loops (Fig. 1) and γH2AX foci (Fig. 4 *A* and *E*) in U2OS cells, the results show that G4 ligands can have specific molecular effects. Although the findings establish that DSB triggered by PDS and FG is mediated by R loops to a large extent, the described differences suggest that the mechanisms of

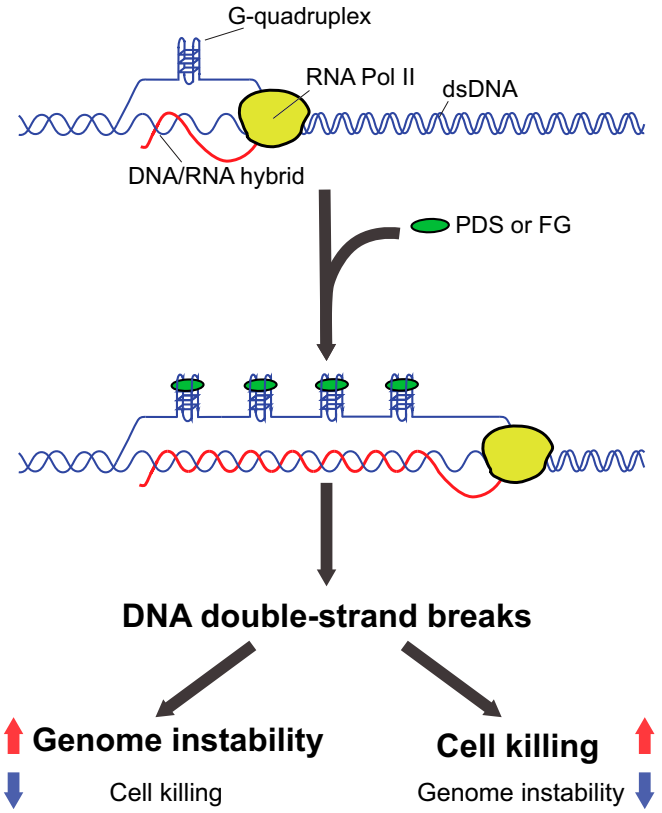

Fig. 6. Molecular model of PDS and FG activity in cancer cells. Ligand-stabilized G4s can cause R-loop spreading at transcribed genes, which results in the accumulation of DNA DSB. DNA breaks can activate molecular pathways leading to either cell killing or micronucleus generation (genome instability).

DNA damage and/or repair may be, at least partially, different. One possibility is that HRR and other DNA repair mechanisms are activated with different strengths and/or chromatin localization of DSB is different for the two studied ligands. Interestingly, active transcription and histone modifications were proposed to regulate DSB repair pathway choice (43, 44). In addition, we cannot completely exclude more selective R loop-independent mechanisms of DNA damage induced by G4 ligands. For instance, the stabilization of G4s in template DNA strands may arrest DNA polymerases triggering replication stress and DNA damage, which may then be resolved by distinct molecular pathways (45). Therefore, it will be interesting to establish in future studies the specificity of DNA damage and repair pathways activated by diverse G4 ligands in relation to stabilized G4s and unscheduled R loops at functionally distinct chromatin sites.

Cooperative interactions between G4s and R loops were previously proposed to occur in *E. coli* and *S. cerevisiae* (15, 36, 46). In particular, the genome instability of a G-rich murine Ig Sμ sequence in yeast was shown to be due to simultaneous formation of G4/R-loop structures under high levels of transcription (36, 46). High transcription levels of the Ig Sμ sequence are required for murine class-switch recombination of Ig genes to likely allow noncanonical DNA structures to form (47). In our study, we found that G4 ligands could trigger cell killing and genome instability with different efficiency. FG was more cytotoxic than PDS, whereas PDS consistently induced micronuclei to a greater extent than FG, particularly in *BRCA2*-depleted cancer cells, with a mechanism involving unscheduled R-loop/G4s and DSB formation (Fig. 6). Micronucleus formation depends on a failure of proper chromosomal DSB repair and requires cell passage through mitosis (35, 37, 38). Interestingly, micronuclei can be a source of cytoplasmic genomic DNA that can activate the STING (stimulator of interferon genes) proinflammation response, eventually leading to activation of the innate immune system (37, 38, 48, 49). Of note, a high frequency of micronucleation was reported in mouse embryonic fibroblasts lacking RNaseH2, a model of monogenic autoinflammation diseases (37, 50). RNaseH2 is an RNaseH enzyme present in mammalian cells, which is involved in ribonucleotide excision repair (51, 52) and can also resolve R loops (11–14). Thus, unscheduled R loops may trigger micronucleus generation in different cell types, and our findings suggest that micronuclei induced by PDS, and to a lesser extent by FG, might lead to an immunostimulatory response in human cancer cells.

Therefore, we have uncovered an R loop-dependent mechanism of DSB accumulation and genome instability caused by the studied G4 ligands in human cancer cells. The mechanistic role played by unscheduled R loops/G4s in the ligand activity can be exploited to discover new anticancer compounds. In addition, our findings foresee the potential of anticancer therapies based on the combination of immunotherapy with G4-targeting small molecules able to elicit an effective innate immune response.

Methods

Compounds. FG and FA were synthesized as described previously (53); IR, ^{1}H NMR, ^{13}C NMR, mass spectral data, and elemental analyses are reported in *SI Appendix, Methods*. Pyridostatin and Braco-19 were purchased from Merck. Chemical reagents were from Merck if not otherwise indicated and were used as indicated in *SI Appendix, Methods*.

Cell Lines. The human U2OS cell line was purchased from ATCC (LGC Standards). Human WI-38 fibroblasts, immortalized with hTERT (54), were kindly obtained from C. Mann (CEA, Gif-sur-Yvette, France) and E. Nicolas (Université de Toulouse). U2OS_T-Rex_RH (expressing an mCherry-tagged RNaseH1) and U2OS_T-Rex cell lines were a kind gift from P. Calsou (IPBS, Toulouse, France), as described already (32). We generated the human U2OS_RH cell line as follows: U2OS cells were first transfected with a pLVX-EF1α-Tet3G-Hygro Tet transactivator-expressing vector and selected with 500 μg/mL hygromycin B. Then, hygromycin-resistant cells were transfected with a pLVX-Tight-Puro vector expressing a FLAG-tagged truncated version of human RNaseH1 (pLVX-Tight-Puro-RH-Flag) and selected with 1.5 μg/mL puromycin. Plasmid vectors were kindly obtained from K. Cimprich (Stanford University, Stanford, CA) (55). All cell lines were routinely tested for mycoplasma (Sigma-Aldrich; MP0035), and cell identity was confirmed with an STR (short tandem repeat) assay at the start and end of the experimental work by BMR Genomics. Cell-culture conditions, cell treatments, and *BRCA2* gene silencing are described in *SI Appendix, Methods*.

Immunofluorescence Microscopy. Slides were visualized at room temperature by using a fluorescence microscope (Eclipse 90i; Nikon) or high-content imaging system (Operetta; PerkinElmer). Cell seeding was performed on a 35-mm dish, 4-well Nunc Lab-Tek II Chamber Slide System (Nalge Nunc; 154526), or 96-well plate (CellCarrier; PerkinElmer) for Operetta massive cell analysis. Plates were coated or not with poly-L-lysine solution (Merck; P4707). After 24 h from seeding, cells were treated with 10 μM PDS, FG, or FA or 15 μM Braco-19 for the indicated time. For high-throughput cell-image analysis, 96-well plates were scanned using the Operetta High-Content Imaging System (Harmony Imaging 4.1; PerkinElmer). After data acquisition, nuclear foci detection and subsequent analyses were performed with Columbus 2.5.0 software (PerkinElmer). For graphical representation of focus distribution, we used box-and-whisker plots using GraphPad Prism 6 software with the following settings: boxes: 25 to 75 percentile range; whiskers: 10 to 90 percentile range; horizontal bars: median number of foci; "+": mean number of foci. Purification and validation of S9.6 and BG4 antibodies are reported in *SI Appendix, Methods*. Detailed protocols of cell fixation and staining for each antibody and cell-image analyses are reported in *SI Appendix, Methods*.

Genome R-Loop Mapping. We used DNA:RNA immunoprecipitation methodologies to immunoprecipitate and isolate DNA:RNA duplexes from genomic DNA preparations by using S9.6 antibody and to map genome-wide R-loop structures, as described previously (18, 27). A detailed DRIP protocol is

reported in *SI Appendix, Methods*. RNA-seq protocols and bioinformatic tools and procedures of genomic R-loop maps are reported in *SI Appendix, Methods*.

Other Methods and Data Availability. Other standard methods [Western blots, cytofluorimetry, MTT (3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide) proliferation assay, quantitative PCR] and primer sequences are reported in *SI Appendix, Methods*. Sequence DRIP reads are available at the Gene Expression Omnibus database (56).

ACKNOWLEDGMENTS. We acknowledge the Toulouse Regional Imaging–Light Imaging Toulouse CBI Platform and Microscopy Core Facility at Medical Faculty, University of Bonn for technical assistance in high-content microscopy. We thank S. Gabor for providing the HB-8730 hybridoma, S. Balasubramanian for the BG4-expressing plasmid, P. Calsou and S. Britton for the mCherry-RNaseH1–expressing cell line, K. Cimprich for the FLAG-RNaseH1–expressing vector, and Barbara Taurisano for technical help. This work was supported by Associazione Italiana per la Ricerca sul Cancro (AIRC) Grant IG-15886 (to G.C.). A.D.M. and M.R. were supported by University of Bologna PhD student fellowships; S.G.M. was supported by a FIRC postdoc fellowship.

1. Hänsel-Hertsch R, Di Antonio M, Balasubramanian S (2017) DNA G-quadruplexes in the human genome: Detection, functions and therapeutic potential. *Nat Rev Mol Cell Biol* 18:279–284.
2. Balasubramanian S, Hurley LH, Neidle S (2011) Targeting G-quadruplexes in gene promoters: A novel anticancer strategy? *Nat Rev Drug Discov* 10:261–275.
3. Cimino-Reale G, Zaffaroni N, Folini M (2016) Emerging role of G-quadruplex DNA as target in anticancer therapy. *Curr Pharm Des* 22:6612–6624.
4. Müller S, et al. (2012) Pyridostatin analogues promote telomere dysfunction and long-term growth inhibition in human cancer cells. *Org Biomol Chem* 10:6537–6546.
5. Rodriguez R, et al. (2012) Small-molecule-induced DNA damage identifies alternative DNA structures in human genes. *Nat Chem Biol* 8:301–310.
6. Zimmer J, et al. (2016) Targeting BRCA1 and BRCA2 deficiencies with G-quadruplex-interacting compounds. *Mol Cell* 61:449–460.
7. Xu H, et al. (2017) CX-5461 is a DNA G-quadruplex stabilizer with selective lethality in BRCA1/2 deficient tumours. *Nat Commun* 8:14432.
8. Ribeyre C, et al. (2009) The yeast Pif1 helicase prevents genomic instability caused by G-quadruplex-forming CEB1 sequences in vivo. *PLoS Genet* 5:e1000475.
9. Vannier J-BB, Pavicic-Kaltenbrunner V, Petalcorin MIR, Ding H, Boulton SJ (2012) RTEL1 dismantles T loops and counteracts telomeric G4-DNA to maintain telomere integrity. *Cell* 149:795–806.
10. van Wietmarschen N, et al. (2018) BLM helicase suppresses recombination at G-quadruplex motifs in transcribed genes. *Nat Commun* 9:271.
11. Santos-Pereira JM, Aguilera A (2015) R loops: New modulators of genome dynamics and function. *Nat Rev Genet* 16:583–597.
12. Sollier J, Cimprich KA (2015) Breaking bad: R-loops and genome integrity. *Trends Cell Biol* 25:514–522.
13. Skourti-Stathaki K, Proudfoot NJ (2014) A double-edged sword: R loops as threats to genome integrity and powerful regulators of gene expression. *Genes Dev* 28:1384–1396.
14. Chédin F (2016) Nascent connections: R-loops and chromatin patterning. *Trends Genet* 32:828–838.
15. Duquette ML, Handa P, Vincent JA, Taylor AF, Maizels N (2004) Intracellular transcription of G-rich DNAs induces formation of G-loops, novel structures containing G4 DNA. *Genes Dev* 18:1618–1629.
16. Brooks TA, Hurley LH (2009) The role of supercoiling in transcriptional control of MYC and its importance in molecular therapeutics. *Nat Rev Cancer* 9:849–861.
17. Kouzine F, et al. (2013) Transcription-dependent dynamic supercoiling is a short-range genomic force. *Nat Struct Mol Biol* 20:396–403.
18. Ginno PA, Lim YW, Lott PL, Korf I, Chédin F (2013) GC skew at the 5′ and 3′ ends of human genes links R-loop formation to epigenetic regulation and transcription termination. *Genome Res* 23:1590–1600.
19. Aguilera A, Gómez-González B (2017) DNA-RNA hybrids: The risks of DNA breakage during transcription. *Nat Struct Mol Biol* 24:439–443.
20. Sparapani S, et al. (2010) Bis-guanylhydrazone diimidazo[1,2-a:1,2-c]pyrimidine as a novel and specific G-quadruplex binding motif. *Chem Commun (Camb)* 46:5680–5682.
21. Amato J, et al. (2016) Toward the development of specific G-quadruplex binders: Synthesis, biophysical, and biological studies of new hydrazone derivatives. *J Med Chem* 59:5706–5720.
22. Marinello J, Chillemi G, Bueno S, Manzo SG, Capranico G (2013) Antisense transcripts enhanced by camptothecin at divergent CpG-island promoters associated with bursts of topoisomerase I-DNA cleavage complex and R-loop formation. *Nucleic Acids Res* 41:10110–10123.
23. Mouly L, et al. (2018) PARP-1-dependent RND1 transcription induced by topoisomerase I cleavage complexes confers cellular resistance to camptothecin. *Cell Death Dis* 9:931.
24. Sanz LA, et al. (2016) Prevalent, dynamic, and conserved R-loop structures associate with specific epigenomic signatures in mammals. *Mol Cell* 63:167–178.
25. Biffi G, Tannahill D, Miller J, Howat WJ, Balasubramanian S (2014) Elevated levels of G-quadruplex formation in human stomach and liver cancer tissues. *PLoS One* 9:e102711.
26. Chambers VS, et al. (2015) High-throughput sequencing of DNA G-quadruplex structures in the human genome. *Nat Biotechnol* 33:877–881.
27. Manzo SG, et al. (2018) DNA topoisomerase I differentially modulates R-loops across the human genome. *Genome Biol* 19:100.
28. Aguilera A, García-Muse T (2012) R loops: From transcription byproducts to threats to genome stability. *Mol Cell* 46:115–124.
29. Skourti-Stathaki K, Proudfoot NJ, Gromak N (2011) Human senataxin resolves RNA/DNA hybrids formed at transcriptional pause sites to promote Xrn2-dependent termination. *Mol Cell* 42:794–805.
30. Nadel J, et al. (2015) RNA:DNA hybrids in the human genome have distinctive nucleotide characteristics, chromatin composition, and transcriptional relationships. *Epigenetics Chromatin* 8:46.
31. Huppert JL, Balasubramanian S (2007) G-quadruplexes in promoters throughout the human genome. *Nucleic Acids Res* 35:406–413.
32. Britton S, et al. (2014) DNA damage triggers SAF-A and RNA biogenesis factors exclusion from chromatin coupled to R-loops removal. *Nucleic Acids Res* 42:9047–9062.
33. Moynahan ME, Jasin M (2010) Cells have evolved various strategies to contend with the multitude of DNA lesions, including DNA strand breaks, that the genome incurs on a continuous basis. *Nat Rev Mol Cell Biol* 11:196–207.
34. Bell JC, Kowalczykowski SC (2016) Mechanics and single-molecule interrogation of DNA recombination. *Annu Rev Biochem* 85:193–226.
35. Hatch EM, Fischer AH, Deerinck TJ, Hetzer MW (2013) Catastrophic nuclear envelope collapse in cancer cell micronuclei. *Cell* 154:47–60.
36. Yadav P, et al. (2014) Topoisomerase I plays a critical role in suppressing genome instability at a highly transcribed G-quadruplex-forming sequence. *PLoS Genet* 10:e1004839.
37. Mackenzie KJ, et al. (2017) cGAS surveillance of micronuclei links genome instability to innate immunity. *Nature* 548:461–465.
38. Harding SM, et al. (2017) Mitotic progression following DNA damage enables pattern recognition within micronuclei. *Nature* 548:466–470.
39. Bartsch K, et al. (2017) Absence of RNase H2 triggers generation of immunogenic micronuclei removed by autophagy. *Hum Mol Genet* 26:3960–3972.
40. García-Muse T, Aguilera A (2016) Transcription-replication conflicts: How they occur and how they are resolved. *Nat Rev Mol Cell Biol* 17:553–563.
41. Maizels N (2005) Immunoglobulin gene diversification. *Annu Rev Genet* 39:23–46.
42. Hänsel-Hertsch R, et al. (2016) G-quadruplex structures mark human regulatory chromatin. *Nat Genet* 48:1267–1272.
43. Marnef A, Cohen S, Legube G (2017) Transcription-coupled DNA double-strand break repair: Active genes need special care. *J Mol Biol* 429:1277–1288.
44. Clouaire T, Legube G (2015) DNA double strand break repair pathway choice: A chromatin based decision? *Nucleus* 6:107–113.
45. Schiavone D, et al. (2016) PrimPol is required for replicative tolerance of G quadruplexes in vertebrate cells. *Mol Cell* 61:161–169.
46. Yadav P, Owiti N, Kim N (2016) The role of topoisomerase I in suppressing genome instability associated with a highly transcribed guanine-rich sequence is not restricted to preventing RNA:DNA hybrid accumulation. *Nucleic Acids Res* 44:718–729.
47. Lee CG, et al. (2001) Quantitative regulation of class switch recombination by switch region transcription. *J Exp Med* 194:365–374.
48. Barber GN (2015) STING: Infection, inflammation and cancer. *Nat Rev Immunol* 15:760–770.
49. Mouw KW, Goldberg MS, Konstantinopoulos PA, D'Andrea AD (2017) DNA damage and repair biomarkers of immunotherapy response. *Cancer Discov* 7:675–693.
50. Mackenzie KJ, et al. (2016) Ribonuclease H2 mutations induce a cGAS/STING-dependent innate immune response. *EMBO J* 35:831–844.
51. Hiller B, et al. (2012) Mammalian RNase H2 removes ribonucleotides from DNA to maintain genome integrity. *J Exp Med* 209:1419–1426.
52. Sparks JL, et al. (2012) RNase H2-initiated ribonucleotide excision repair. *Mol Cell* 47:980–986.
53. Andreani A, et al. (2004) Potential antitumor agents. 34.(1) Synthesis and antitumor activity of guanylhydrazones from imidazo[2,1-b]thiazoles and from diimidazo[1,2-a:1,2-c] pyrimidine. *Anticancer Res* 24:203–211.
54. Jeanblanc M, et al. (2012) Parallel pathways in RAF-induced senescence and conditions for its reversion. *Oncogene* 31:3072–3085.
55. Stork CT, et al. (2016) Co-transcriptional R-loops are the main cause of estrogen-induced DNA damage. *eLife* 5:e17548.
56. Russo M, et al. (2018) R-loop maps in U2OS cells after G-quadruplex ligands treatment. Gene Expression Omnibus. Available at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE115957. Deposited June 18, 2018.

ANNEX 2

Russo et al. BMC Bioinformatics (2019) 20:414
https://doi.org/10.1186/s12859-019-3009-9

BMC Bioinformatics

SOFTWARE Open Access

DROPA: DRIP-seq optimized peak annotator

Marco Russo[1], Bruno De Lucca[1], Tiziano Flati[2,3], Silvia Gioiosa[2,3], Giovanni Chillemi[2,4] and Giovanni Capranico[1*]

Abstract

Background: R-loops are three-stranded nucleic acid structures that usually form during transcription and that may lead to gene regulation or genome instability. DRIP (DNA:RNA Immunoprecipitation)-seq techniques are widely used to map R-loops genome-wide providing insights into R-loop biology. However, annotation of DRIP-seq peaks to genes can be a tricky step, due to the lack of strand information when using the common basic DRIP technique.

Results: Here, we introduce DRIP-seq Optimized Peak Annotator (DROPA), a new tool for gene annotation of R-loop peaks based on gene expression information. DROPA allows a full customization of annotation options, ranging from the choice of reference datasets to gene feature definitions. DROPA allows to assign R-loop peaks to the DNA template strand in gene body with a false positive rate of less than 7%. A comparison of DROPA performance with three widely used annotation tools show that it identifies less false positive annotations than the others.

Conclusions: DROPA is a fully customizable peak-annotation tool optimized for co-transcriptional DRIP-seq peaks, which allows a finest gene annotation based on gene expression information. Its output can easily be integrated into pipelines to perform downstream analyses, while useful and informative summary plots and statistical enrichment tests can be produced.

Keywords: R-loop, Non-canonical DNA structures, Genome annotation, Next-generation sequencing

Background

R-loops are three stranded nucleic acid structures composed by a DNA:RNA hybrid duplex and a displaced ssDNA (single strand DNA) strand. R-loops form co-transcriptionally when nascent RNAs anneal back to DNA template strand [1, 2]. R-loops have been shown to be involved in many nuclear processes such as transcription regulation, DNA methylation modulation and DNA repair mechanisms. However unscheduled R-loop formation is associated with DNA damage accumulation, genome instability and genetic diseases [3].

Genome-wide maps of these peculiar nucleic acid structures have boosted our understanding of R-loop biology [1]. Immunoprecipitation-based techniques, generally known as DRIP (DNA:RNA Immunoprecipitation), coupled with parallel sequencing (DRIP-seq), are widely used to maps R-loops genome-wide [4, 5] and several DRIP variants have been developed with the intent to improve the identification of genomic R-loops [3]. However, the most common technique (DRIP) allows the detection of R-loop regions without providing the strand information of the DNA:RNA hybrid. Understanding the DNA strand forming the hybrid is essential to investigate the dynamic interplay of R-loops with other nucleic acid structure (e.g. G-quadruplexes) [6] or with basic directional mechanisms such as replication and transcription [7].

Moreover, DRIP-seq data are commonly analyzed with standard peak callers, such as MACS (Model-based Analysis of ChIP-Seq) [8], to identify regions with above-threshold coverage signals, usually called “peaks”. Nevertheless, DRIP peaks are markedly different from traditional ChIP (Chromatin Immunoprecipitation) peaks of transcription factors as the former peaks are usually much longer than the latter ones, spanning across several genes features or different genes. As R-loops can span several gene features (REFs), the assignment of R-loop peaks to a unique feature may not be appropriate.

To overcome these issues, we have developed a new software, named DROPA (DRIP-seq Optimized Peak Annotator), which makes use of gene expression data to annotate R-loop peaks to strand templates and expressed

* Correspondence: giovanni.capranico@unibo.it
[1]Department of Pharmacy and Biotechnology, University of Bologna, Bologna, Italy
Full list of author information is available at the end of the article

genes. Thus, DROPA allows the identification of the DNA strand annealed to the RNA and annotates R-loop.

Implementation

Program architecture and design

DROPA is a command-line tool, developed in Python, and it can be launched in Unix environment (Linux, MacOS, Windows Subsystem for Linux).

DROPA consists of six Python scripts, *PeakOverlap*, *CheckExpression*, *FeatureAssign*, *TableCreator*, *RandPeak*, and *SummaryPlot* (Fig. 1).

- *PeakOverlap* searches for genes overlapping R-loop peaks. Two BED (Browser Extensible Data) files are produced as output: one lists peaks with corresponding overlapping genes and the other lists intergenic peaks without overlapping genes. Notwithstanding there are some libraries in R (e.g. GenomicRanges [9]) that can perform this step, however we wrote *PeakOverlap* in Python to be consistent with the next scripts.
- *CheckExpression* introduces the main novelty of DROPA as compared with common peak annotation tools, as it considers gene expression levels in order to assign each R-loop peak to a given gene. It considers all overlapping genes of a peak and, if only one gene overlaps with the query peak, then that gene is assigned to the peak. In case of multiple genes overlapping to the same peak, *CheckExpression* evaluates their transcription levels and selects the gene with the highest level. Background levels can be set providing a threshold. If expression levels are below thresholds, they are the same for all overlapping genes, or if expression data are not provided, then the function selects the gene with the largest overlap with the query peak. Gene expression data can be in TPM (Transcripts Per Million), FPKM (Fragments Per Kilobase Million) or any other normalized values.
- *FeatureAssign* identifies all gene regions (upstream/downstream region, intron, exon, UTR (Untranslated Region) regions) overlapping to the peak.
- *TableCreator* returns a table that reports relevant information of annotated genes (name, template strand and other features) for each peak.
- *RandPeak* (optional) performs analyses of random R-loop peaks to calculate gene feature enrichment scores. The script takes the query peaks coordinates and returns randomly shuffled peaks all over the genome using BEDtools shuffle tool [10]. Then, it launches the 1 to 4 scripts of DROPA analyses for the random peaks.
- Once all the steps are performed, *SummaryPlot* is used to plot results.

Input data

DROPA requires three input data, which are:

- A file containing query peak locations in BED format;
- A reference set folder containing information about genes features (5'UTR, 3'UTR, exon, intron) in BED

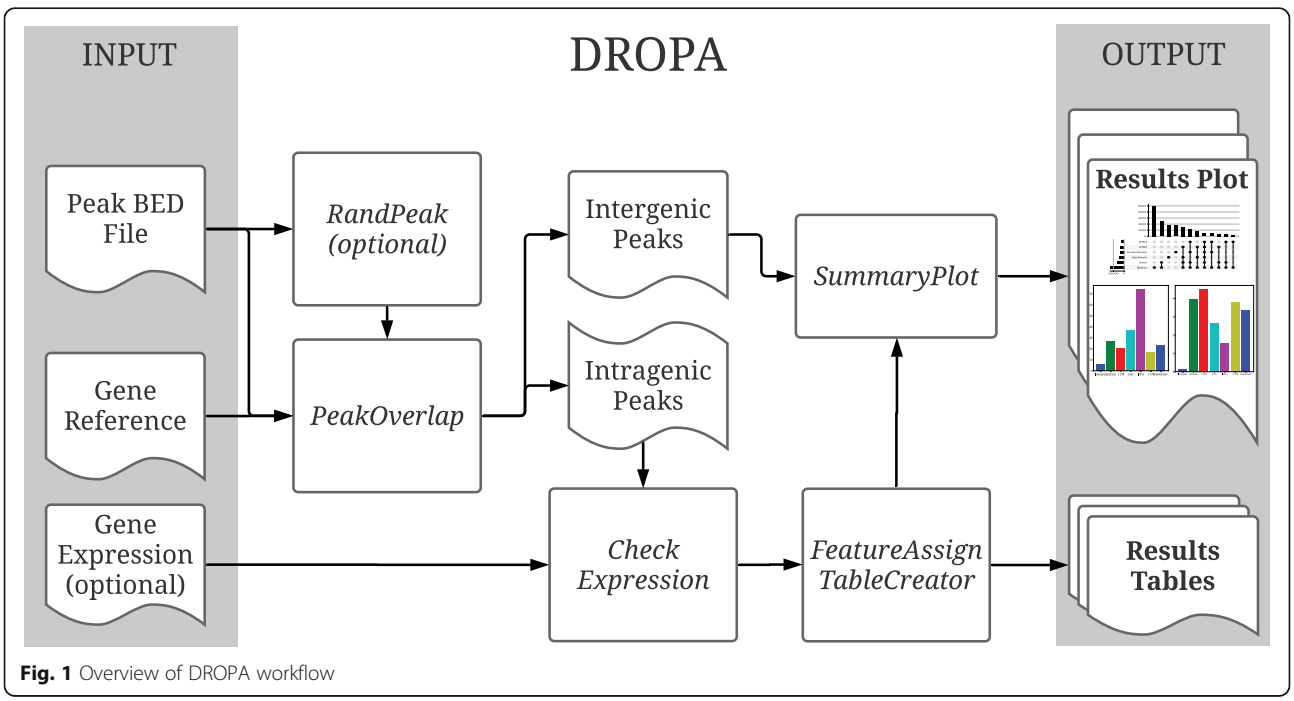

Fig. 1 Overview of DROPA workflow

format and a gene reference in BED12. We provide many ready-to-use reference set for *Homo sapiens* (hg19) and for *Mus musculus* (mm9 and mm10). Gene reference can be easily generated for every genome of interest, and also custom gene reference can be used.
- A 2-column gene expression table containing the name of each gene and its normalized expression value (FPKM, TPM, etc.). This table can be optional.

Besides the three input data, other custom parameters can be provided:

- The size of upstream and downstream regions.
- The gene expression threshold to consider a gene as expressed.
- The number of shuffle samples to perform the randomization analysis.

Output data format

DROPA output consists of a folder containing table and image files.

The main table file result is the "*annotation.table*" that contains, for every annotated peak:

- Peak coordinates: chromosome, peak start, peak end;
- Peak name, as in the input file;
- Name of the gene, his strand and his expression value, as in the reference;
- Which features of the gene are covered by the peak (Upstream, 5'UTR, Exon, Intron, 3'UTR, Downstream).
- A warning flag if the peak is localized in a region in which antisense R-loops can form.

DROPA provides other secondary table files, such as the list of intergenic peaks and summary tables used to create plot figures. Figures and the annotation.table are provided in three version: the "expressed" in which are reported results for peaks annotated to genes with expression value above the threshold, the "unexpressed" in which are reported the ones annotated to genes with expression value below the threshold, and the "merged" which reports the aggregation of the previous two.

DROPA produces many informative summary plots, regarding the percentage of peaks overlapping each genic feature (Fig. 2a), or their proportion as a pie chart (Fig. 2b). Furthermore, since many peaks usually overlap more than one feature, DROPA provides a plot in which is shown the number of peaks that overlap each combination of feature (Fig. 2c). Finally, if enrichment analysis is performed, it is provided a histogram (Fig. 2d) with standard deviations bars and *p*-value of a chi-squared contingency test, showing the fold enrichment for each gene feature, calculated as the ratio between the number of query peaks that overlap a feature and the mean number of randomly shuffled peaks.

Availability

DROPA package can be downloaded from https://github.com/marcrusso/DROPA.

Installation

Detailed installation guide for DROPA and all python libraries required is available at https://github.com/marcrusso/DROPA. (see also Additional file 1 for DROPA requirements)

Launching

To launch DROPA with default settings this command can be used:

```
python3 DROPA_v1.0.0.py -ref GeneReference/GeneReferenceSet/ -o OutputFolderName QueryPeak.bed
```

Results

Influence of expression data metrics on DROPA

As the main feature of DROPA is peak annotation using expression levels, we tested whether different expression metrics (TPM and FPKM) lead to different annotation output. In this analysis, the same default settings were used. The comparison showed that using TPMs more peaks (212, 1.3%) were assigned to expressed genes. However, all other peaks (15,872, 98.7%) were assigned to the same gene using FPKM or TPM values (see Additional file 1: Table S1). Overall the results show that using TPM or FPKM substantially leads to very similar overall peak annotation.

Assessment of DROPA performance

To assess the correct annotation rate of DROPA, we determined the correct assignment of a query dataset of DRIPc-seq peaks [11] to the DNA template strand. DRIPc is a DRIP technique variant that maintains the strand information of DNA:RNA hybrid peaks. Our comparison shows that, when DROPA assigns peaks based on gene expression (76,526 peaks, see Additional file 1: Table S2), 88.6% (67,796) of them are assigned correctly (see Additional file 1: Table S3). Among the 11.4% (8730) of peaks with wrong annotation, we noticed that 5.05% (3871) are in the same position of another DRIPc peaks but in the opposite strand, and another 3.53% (2707) are mapped within 5000 bp upstream or downstream to expressed genes (see Additional file 1: Figure S1). As antisense transcription is known in particular at 5′ and 3′ ends of expressed genes, these analyses suggest that many peaks assigned to the wrong strand are potential antisense R-loops [11]. Therefore, if we consider only the transcribed regions of a gene, DROPA efficiency is 93.8% (see Additional file 1:

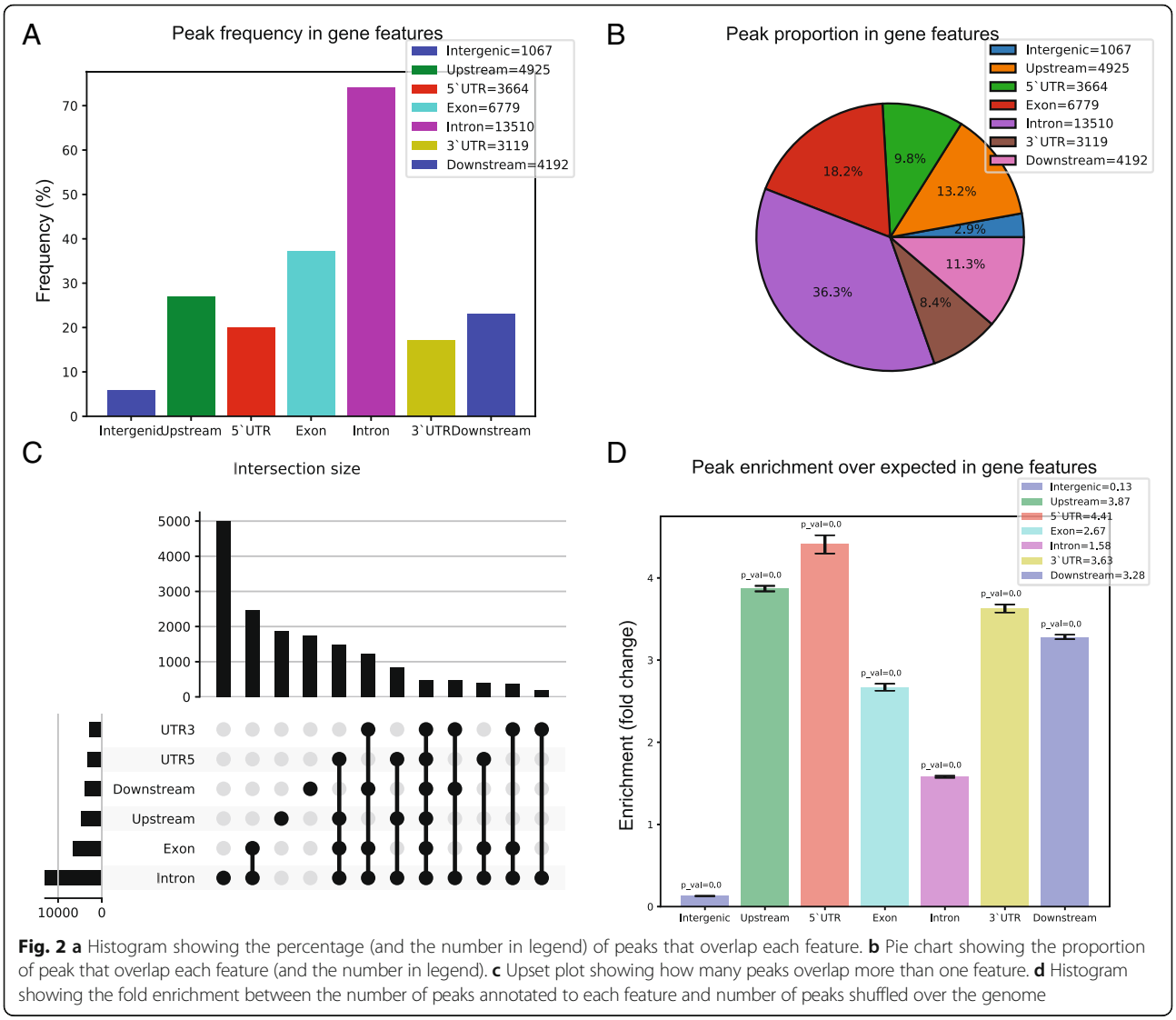

Fig. 2 **a** Histogram showing the percentage (and the number in legend) of peaks that overlap each feature. **b** Pie chart showing the proportion of peak that overlap each feature (and the number in legend). **c** Upset plot showing how many peaks overlap more than one feature. **d** Histogram showing the fold enrichment between the number of peaks annotated to each feature and number of peaks shuffled over the genome

Figure S1). In order to warn the user about peaks that can be ambiguously assigned as they map in regions where antisense R-loop can form, we provide: i) specific output tables of these peaks and ii) warning flags to these peaks in the main output tables, leaving therefore the user the option to exclude them for further analysis.

DROPA comparison with existing tools

To assess DROPA performance, we compared it with three widely used or recent annotation tools: HOMER [12], PAVIS [13] and UROPA [14], which are based on different algorithms. PAVIS and HOMER annotate peaks based on the nearest TSS (Transcription Start Site), while UROPA allows to choose between the nearest start, end or center of the reference region. DROPA annotates all gene features (UTR regions, exons, introns, etc.) overlapping to peaks, while HOMER and PAVIS select one gene feature only. This may limit the biologically relevant information of HOMER and PAVIS output data when query peaks have a size larger than the gene features.

DROPA is highly flexible and customizable. DROPA, PAVIS and Homer have default gene reference sets that make these tools ready-to-use: DROPA has human (hg19) and mouse (mm9, mm10) genome as default, while PAVIS has gene sets of many organisms and genome assemblies. However, DROPA allows the choice of any custom gene set. Among the others, only PAVIS does not allow the use of a custom gene reference set. HOMER does not allow to set the size of upstream and downstream gene regions, which is useful while working on peaks that can form kilobases far from a gene, and it is instead customizable with DROPA, PAVIS and UROPA.

DROPA lacks a Graphical User Interface (GUI), however it is easy to use thanks to few and fully described command flags, which make it easily integrable into pipelines. PAVIS, a web-based tool, offers a GUI and requires

an Internet connection, while UROPA offers both a command-line tool and a web-based GUI. HOMER is available only as command-line tool.

Although all four tools produce a full annotation table for every query peak, DROPA also produces summary tables and many other plots of peak distribution or enrichment over specific gene features. PAVIS produces a summary table and a pie chart of peak annotations, HOMER provides no plots but only a summary table, while UROPA also produces a summary report with plots.

Feature comparison summary is reported in Table 1.

We compared the four tools output using an experimental set of R-loops peaks determined by DRIP-seq in human cells [6]. Peak calling was performed using MACS and comparisons were carried out using the same gene reference and the same upstream/downstream dimension. Briefly, DROPA gives different annotation results in comparison with all three tools in analysis. A full description of annotation result is reported in Additional file 1.

Intergenic peaks in DROPA data are always fewer compared to HOMER (10.8% of total query peaks versus 17.9% in HOMER), PAVIS (5.7% versus 12.1%) and UROPA (3.3% versus 11.2%). This is mainly due to the fact that DROPA does not take in account only the center of the query peak (that for peaks that have a dimension of kilobases can be far from the gene region) for annotation, but both the start and end point. About 10, 20 and 22% of query peaks are annotated with different genes by DROPA with respect to HOMER, PAVIS and UROPA, respectively.

As R-loop formation is mainly a co-transcriptional phenomenon and PAVIS and HOMER primarily rely on closest TSS, we can argue that DROPA identifies less false positive annotations as compared to PAVIS and HOMER due to the use of expression data (see Additional file 1: Figures S2 and S3). Although UROPA annotation does not rely on closest TSS search but rather on overlap, its annotation approach still gives a result clearly different from DROPA one (see Additional file 1: Figure S4), which, is optimized using gene expression data.

Limitations

Even though DROPA can define and assign co-transcriptional peaks in the body of expressed genes with good efficiency, its main limitation is the detection and correct assignment of antisense R-loop peaks. To compensate for this limitation, DROPA provides a list of peaks assigned only to the upstream/downstream region of expressed gene, where antisense transcripts can be present. Moreover, in the main annotation table a warning flag indicates either peaks in upstream/downstream regions and peaks located at overlapping expressed genes. A comparison of this information with genomic datasets of antisense transcripts can help the user for further analysis.

Conclusions

DROPA is a full customizable peak annotation tool optimized for co-transcriptional DRIP-seq peaks, allowing a finest gene annotation based on gene expression information. Since the expression data table is optional, this tool can be used with other sequencing data regarding genomic features that are not strictly associated with TSS (for which tools like PAVIS and HOMER are developed) and that are characterized by broad peak dimension, such as Histone marks IP-seq, DNAse-seq and FAIRE-seq. Using DROPA, users can take advantage from its alternative annotation algorithm, based on largest overlap with the query peak, the multi-feature annotation and the informative summary plots.

Methods

In the following evaluation, DROPA was tested on a machine running Ubuntu OS (vers. 16.04 LTS) with 8 CPU cores and 16 GB of RAM.

Influence of expression data metrics on DROPA

To perform evaluation of DROPA results using different gene expression values we used an experimental set of DRIP-seq peak (available at GEO: GSE115957) and his relative RNA-seq data. Using Stringtie (REF), TPM and FPKM were computed using RNA-seq data using RefSeq gene reference. Peak dataset and gene expression table are available in DROPA repository as Test_hg19_DRIP_

Table 1 Feature comparison between DROPA and PAVIS, HOMER and UROPA

	DROPA	PAVIS	HOMER	UROPA
Offline Usage	✓	✗	✓	✓
Pipeline integration	✓	✗	✓	✓
Reference gene set customization	✓	✗	✓	✓
Upstream/downstream region definition	✓	✓	✗	✓
Multiple gene feature annotation	✓	✗	✗	✗
Statistical enrichment over gene feature	✓	✓	✓	✗
Summary plot Results	✓	✓	✗	✓

peaks.bed. DROPA was launched two times with default settings using TPM or FPKM table as expression data and hg19_Refseq as gene reference. Results of annotation were compared counting how many peaks were annotated as intergenic and how many peaks were annotated to the same gene.

Assessment of DROPA performance

To perform evaluation of DROPA using stranded data we downloaded a DRIPc-seq peak dataset and relative RNA-seq data (available at GEO: GSE70189). Peak dataset and gene expression table are available in DROPA repository. DROPA was launched with default settings using as input the peak dataset, the gene expression table and the gene reference hg19_UCSCgenes. Then strandness of peaks annotated on expressed genes was compared with the one of the original dataset.

DROPA comparison with existing tools

In all 3 comparison we used an experimental set of DRIP-seq peak (available at GEO: GSE115957) and his relative RNA-seq data. Peak dataset and gene expression table are available in DROPA repository as Test_hg19_DRIP_peaks.bed and Test_hg19_RefSeq_Expression. Since each tool has different degree of customization (fixed upstream/downstream dimension, gene reference selection, etc.), we adapted DROPA settings to the one of the tool in analysis. In comparison with HOMER, DROPA was launched with upstream/downstream region dimensions set to 1 kb and RefSeq gene reference. HOMER was launched with default settings. In comparison with PAVIS, DROPA was launched with default settings and UCSCknown gene reference, while PAVIS was launched setting the Upstream/Downstream region to 5 kb (same as DROPA default). In comparison with UROPA, DROPA was launched with default settings and Ensembl gene reference, while UROPA was launched setting the Upstream/Downstream region to 5 kb. In all three comparison, after peak annotation, results were compared counting how many peaks were annotated as intergenic and how many peaks were annotated to the same gene.

Availability and requirements

Project name: DROPA

Project home page: Source code on https://github.com/marcrusso/DROPA

Operating system: Unix (Linux or Mac OS or Windows Subsystem for Linux)

Programming language: Python3

Other requirements: All Python libraries requirements are listed in Supplementary. Bedtools software is required for peak randomization.

License: MIT license

Any restrictions to use by non-academics: None.

Additional file

Additional file 1: Supplementary file containing DROPA requirements, summary tables and figures ragarding comparison results and a benchmark section. (DOCX 734 kb)

Abbreviations

BED: Browser Extensible Data; ChIP-seq: Chromatin Immunoprecipitation; DRIP: DNA:RNA Immuno Precipitation; DROPA: DRIP Optimized Peak Annotator; FPKM: Fragments Per Kilobase Million; MACS: Model-based Analysis of ChIP-Seq; ssDNA: Single-strand DNA; TPM: Transcripts Per Million; TSS: Transcription Start Site; UTR: Untranslated Region

Acknowledgements

We wish to thank all the members of the laboratory for the helpful discussion.

Authors' contributions

MR and GCa conceived the idea and wrote the manuscript. MR, BDL, TF and SG developed DROPA. MR conducted DROPA comparison with other tools and interpreted the data. GCh revised the tool algorithms. All authors read and approved the final version of the manuscript.

Funding

This work was supported by Associazione Italiana per la Ricerca sul Cancro, Milan [IG15886 to G. Capranico] and University of Bologna PhD fellowship program [to M.R.]. The funding bodies did not influence the design of the study and collection, analysis, and interpretation of data or writing the manuscript.

Availability of data and materials

DROPA repository containing source code, installation guide and usage information is available at https://github.com/marcrusso/DROPA.
The datasets analyzed during the current study are available in the DROPA repository.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Department of Pharmacy and Biotechnology, University of Bologna, Bologna, Italy. [2]National Council of Research, CNR, Institute of Biomembranes, Bioenergetics and Molecular Biotechnologies, Bari, Italy. [3]SCAI-Super Computing Applications and Innovation Department, CINECA, Rome, Italy. [4]Department for Innovation in Biological, Agro-food and Forest systems (DIBAF), University of Tuscia, Viterbo, Italy.

Received: 1 April 2019 Accepted: 29 July 2019
Published online: 06 August 2019

References

1. Chédin F. Nascent connections: R-loops and chromatin patterning. Trends Genet. 2016;32:828–38. https://doi.org/10.1016/j.tig.2016.10.002.
2. Santos-Pereira JM, Aguilera A. R loops: new modulators of genome dynamics and function. Nat Rev Genet. 2015;16:583–97. https://doi.org/10.1038/nrg3961.
3. Crossley MP, Bocek M, Cimprich KA. R-loops as cellular regulators and genomic threats. Mol Cell. 2019;73:398–411. https://doi.org/10.1016/j.molcel.2019.01.024.
4. Ginno PA, Lott PL, Christensen HC, Korf I, Chédin F. R-loop formation is a distinctive characteristic of Unmethylated human CpG Island promoters. Mol Cell. 2012;45:814–25. https://doi.org/10.1016/j.molcel.2012.01.017.

5. Manzo SG, Hartono SR, Sanz LA, Marinello J, De Biasi S, Cossarizza A, et al. DNA topoisomerase I differentially modulates R-loops across the human genome. Genome Biol. 2018;19:100. https://doi.org/10.1186/s13059-018-1478-1.
6. De Magis A, Manzo SG, Russo M, Marinello J, Morigi R, Sordet O, et al. DNA damage and genome instability by G-quadruplex ligands are mediated by R loops in human cancer cells. Proc Natl Acad Sci U S A. 2019;116:816–25. https://doi.org/10.1073/pnas.1810409116.
7. Hamperl S, Bocek MJ, Saldivar JC, Swigut T, Cimprich KA. Transcription-replication conflict orientation modulates R-loop levels and activates distinct DNA damage responses. Cell. 2017;170:774–786.e19. https://doi.org/10.1016/j.cell.2017.07.043.
8. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008;9:R137. https://doi.org/10.1186/gb-2008-9-9-r137.
9. Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, et al. Software for computing and annotating genomic ranges. PLoS Comput Biol. 2013;9:e1003118. https://doi.org/10.1371/journal.pcbi.1003118.
10. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2. https://doi.org/10.1093/bioinformatics/btq033.
11. Sanz LA, Hartono SR, Lim YW, Steyaert S, Rajpurkar A, Ginno PA, et al. Prevalent, dynamic, and conserved R-loop structures associate with specific Epigenomic signatures in mammals. Mol Cell. 2016;63:167–78. https://doi.org/10.1016/j.molcel.2016.05.032.
12. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell. 2010;38:576–89. https://doi.org/10.1016/j.molcel.2010.05.004.
13. Huang W, Loganantharaj R, Schroeder B, Fargo D, Li L. PAVIS: a tool for peak annotation and visualization. Bioinformatics. 2013;29:3097–9. https://doi.org/10.1093/bioinformatics/btt520.
14. Kondili M, Fust A, Preussner J, Kuenne C, Braun T, Looso M. UROPA: a tool for universal RObust peak annotation. Sci Rep. 2017;7:2593. https://doi.org/10.1038/s41598-017-02464-y.

Publisher's Note