UNIVERSITY OF CALIFORNIA
RIVERSIDE

G-quadruplexes in Gene Regulation and Telomere Maintenance

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Environmental Toxicology

by

Jun Yuan

September 2023

Dissertation Committee:
Dr. Yinsheng Wang, Chairperson
Dr. Jikui Song
Dr. Linlin Zhao

The Dissertation of Jun Yuan is approved:

_____

_____

_____

Committee Chairperson

University of California, Riverside

ACKNOWLEDGEMENT

I would like to express my sincere thanks to those who have played a pivotal role in bringing my dissertation to fruition. First and foremost, I would like to express my deepest gratitude to my research advisor, Dr. Yinsheng Wang, whose unwavering support, invaluable insights, and unmatched expertise have been the linchpin of this scholarly journey. He has been the embodiment of patience, especially during those moments when I was overwhelmed by doubts and uncertainties. His passion for research propels me to explore beyond the horizon of conventional thought and encouraging me to move further on academic career.

I am also deeply appreciative of my dissertation committee members, Dr. Jikui Song and Dr. Linlin Zhao. Their patience, steadfast support, insightful criticisms, and invaluable suggestions have been instrumental in elevating the quality of this research.

I am truly humbled to express my gratitude to my senior fellows, especially Dr. Jiabin Wu and Dr. Xiaomei He, whose mentorship, assistance, and unwavering support have been nothing but short of transformative. Their profound expertise and professional experiences in chemical biology experiments have illuminated my PhD journey to today. They offered me a unique perspective of research and the beauty of academic collaboration. I would also like to extend my heartfelt appreciation to our lab manager, Shuli Zhai. Beyond ensuring the seamless daily operations of the lab, Shuli has also been a steadfast pillar of emotional support in both our research endeavors and personal lives.

To my esteemed graduate peers, Dr. Jiekai Yin and Dr. Zi Gao, our shared academic voyage has been a tapestry of both challenges and accomplishments. Our fervent brainstorming, nocturnal dialogues, and the myriad hours spent in the lab have significantly enriched my PhD

research, gifting me with enduring lessons and cherished memories. My gratitude also extends to our lab colleagues, notably Dr. Yuxiang Sun, Dr. Feng Tang, Dr. Ming Huang, Dr. Rong Cai, Dr. Tianyu Qi, Dr. Xiaochuan Liu, and Dr. Yenyu Yang, for their invaluable support and pivotal contributions throughout my doctoral journey.

I would like to express heartfelt appreciation to my dear friends, Dr. Qi Li, Dr. Mengwan Li, Dr. Jian Zhang, Dr. Jiayan Shi, and Dr. Musen Zhou. From lobster fishing excursions to skiing trips in Big Bear, our shared experiences have enriched my PhD journey, elevating it from a mere academic pursuit to a monumental chapter of my life. Our laughter, shared meals, and night conversations were the perfect refuge during the most demanding times. Your companionship reaffirmed my belief in the unparalleled strength of friendship, which stands shoulder to shoulder with familial bonds and romantic ties, shielding us from life's inherent solitude.

Last but not least, I want to express my deepest gratitude to my parents, who always stand behind me as my constant pillars of strength. You celebrated my every milestone, no matter how small, and stood by me during the challenges, reminding me that every stumbling block was but a steppingstone. Words seem inadequate to express the profound gratitude and love I feel as I reflect on the role you have played in shaping both this thesis and my life. Thank you for your unwavering belief, unconditional love, and relentless support!

To everyone acknowledged and the many others who have played a part in this memorable PhD journey, I convey my profound gratitude. This dissertation stands as evidence of the invaluable support you've generously provided through the past days and nights. Thank you from the deepest of my heart.

COPYRIGHT ACKNOWLEDGEMENTS

The text and figures, in part or full, are a reprint of the material as they appear in the following publications:

Chapter 3: Yuan, J., He, X. and Wang, Y. (**2023**) G-quadruplex DNA contributes to RNA polymerase II-mediated 3D chromatin architecture. *Nucleic Acids Res.,* gkad588.

Chapter 4: He, X., Yuan, J. and Wang, Y. (**2021**) G3BP1 binds to guanine quadruplexes in mRNAs to modulate their stabilities. *Nucleic Acids Res.,* 49, 11323-11336.

DEDICATION

To my family and friends!

ABSTRACT OF THE DISSERTATION

G-quadruplexes in Gene Regulation and Telomere Maintenance

by

Jun Yuan

Doctor of Philosophy, Graduate Program in Environmental Toxicology
University of California, Riverside, September 2023
Dr. Yinsheng Wang, Chairperson

The pervasive presence of guanine quadruplex (G4) structures within regulatory regions of the genome has spurred intense research efforts to understand their roles in modulating cellular pathways. Numerous investigations affirm that both DNA and RNA G4s actively partake in pivotal biological processes, including DNA replication, transcription, RNA metabolism, translation, and telomere maintenance. However, further investigation is needed to understand the molecular mechanisms through which these secondary nucleic acid structures modulate biological processes. This dissertation focuses on utilizing bioinformatic analyses of publicly accessible datasets to unravel multifaceted roles of G4 structures in histone modifications, three-dimensional chromatin configurations, RNA metabolism, and telomere maintenance.

In chapter 2, we revealed the G4 co-localization pattern with transcription factors and constructed an interaction network of candidate G4-interacting proteins. Moreover, we explored the interplay between G4 structures and histone marks, unveiling G4 structures as active transcription marks and potential regulators for histone modifications.

In chapter 3, we conducted an intuitive overlapping analysis of previously published RNAPII ChIA-PET and BG4 ChIP-seq data, and our work revealed a strong positive correlation between RNAPII-linked DNA loops and G4 structures in chromatin. In conjunction with HiChIP-seq and RNA-seq, we unveiled vital role of DNA G4 in RNAPII-associated DNA looping and transcription regulation.

In chapter 4, we employed a bioinformatic approach based on the analysis of overlap between RNA G4 (rG4)-seq analysis and eCLIP-seq datasets generated from the ENCODE project. We identified a large number of candidate rG4-binding proteins. We validated that one of these proteins, G3BP1, is a direct binder of rG4 structures, and documented a rG4-dependent function in regulating mRNA stabilities and translation efficiencies.

In chapter 5, I proposed a novel approach for identifying putative telomere-binding proteins through enrichment analysis of ChIP-seq datasets covering zinc finger proteins. Three prominent targets, i.e., ZNF24, ZNF316 and ZBTB33, exhibited significant enrichment with telomeric sequences. A detailed examination of ZBTB33 suggested a potential G4-dependent telomere binding activity of the protein.

In summary, this dissertation introduces a pioneering bioinformatic approach for investigating the intricate interplay between G4 structures and other cellular regulatory mechanisms, identifying new RNA-binding proteins and probing potential G4-dependent telomere-binding proteins. These insights underscore the regulatory significance of G4 structures and shed light on their intricate roles in the nuanced modulation of cellular processes.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1: Introduction

1.1. Guanine-quadruplexes (G4)

1.1.1. G4 formation

The genetic information flows as in central dogma that DNA encodes RNA and RNA encodes protein. In such processes, nucleic acids are essential molecules that not only serve as inheritable materials but also qualitatively and quantitatively regulate this information flow. While RNA adopts versatile secondary structures, DNA primarily consists of its iconic B-form double helix structure (1). Both canonical structure of RNA or DNA are stabilized by Watson-Crick base pairing that adenine pairs with thymine (uracil in RNA) and cytosine pairs with guanine.

The guanine quadruplexes (G4) are four-stranded non-canonical nucleic acid secondary structures folded in guanine-rich regions (2,3). Hoogsteen hydrogen bonding together with coordinated monovalent cation, in the order $K^+ > Na^+ > Li^+$, brings four guanines together to form one layer of guanine quartet structure. Further stacking of at least three layers of G quartets gives rise to G4 structure.

Early models assumed DNA G4s to have loop lengths no longer than seven nucleotides and entail four continuous runs of Gs. Later studies revealed diverse topologies consist of either intramolecular (a single nucleic acid strand folds back on itself) or intermolecular (two or more separate strands bond together), either parallel (all the strands run in the same direction), antiparallel (the strands run in opposite directions), or mixed (the strands run in mixed direction) and even bulges because of discontinuities of guanine track (4).

RNA G-quadruplex exhibits several differences with its DNA counterpart. The presence of 2'-hydroxyl group on the ribose sugar ring enhances stability of RNA G4s since it provides additional intramolecular interactions. The ordered 2'-hydroxyl group also brings water molecules to RNA G4 grooves thus further stabilizes RNA G4s. It also strains the G4 topology as the orientation of the base is strongly favoring the anti-conformation instead of syn-conformation (5). Compared to versatile strand orientations in DNA G4, RNA G4s mainly adopt parallel conformations (6).

1.1.2. G-quadruplex identification and localization

While first reported to be a thermally stable nucleic acid structure in 1962 (7), G-quadruplex was found to be biologically relevant after a discovery made 40 years later where the promoter region of the *c-MYC* oncogene can form G4 structure (8). As in the same year, crystal structure analysis revealed that 21-nucleotide human telomeric sequence can also form a parallel G4 (9). Subsequent studies extended the presence of G4 structures in genomic context including the proximal promoter region of oncogenes, the insulin gene, fragile X syndrome triplet repeats, HIV-1 RNA and telomeric repeat-containing RNA (TERRA) (10). As G4 structures showed significant relevance to biological processes, multiple assays have been developed to detect, map and visualize G4 *in vitro* and *in vivo*.

Biochemical assays provide simple but reproducible and reliable approaches for studying G4-forming sequence *in vitro*. Dimethylsulfate (DMS) footprinting method methylates the guanine residues on single-stranded DNA instead of G4 thus can identify the guanine bases involved in G4 formation (11). Coupling with native polyacrylamide gel, DMS footprinting can further distinguish intramolecular forms of G4 from intermolecular forms. Polymerase stop assay provides an easy and quick way to identify and position nucleic acid secondary structures *in vitro*

as DNA polymerases or reverse transcriptases are incapable of traversing through G4 and will stop before G4 structure (12).

Circular dichroism (CD) spectroscopy, which measures the differential absorption of left- and right-handed circularly polarized light, is a standard biophysical method for characterizing G4s (13). Because the unique CD signature of G4 structure, it has been widely used to assess the effect of small molecule ligand and metal cation on the folding and topology of G4 structures. X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy can offer detailed insights into the precise arrangement of atoms within G4 structure (14,15). While X-ray crystallography requires significant efforts in crystallization and may not capture the structure in biological contexts, NMR can not only determines the atomic-resolution structures but also provide kinetics and dynamics of G4 structures in the physiological state.

Computational prediction using the consensus sequence motif of $G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}N_{1-7}$ identified over 370,000 putative G4 sequences (PQS) in the human genome, where most of them are enriched in genome regulatory regions including telomeres, promoters and 5'-untranslated regions (16). High-throughput sequencing was later ultilized to further localize G4 structure at the genome-wide scale. Polymerase stalling caused by ligand-stabilized G4 structure during sequencing (G4-seq) gives over 700,000 observed G4 sequences (OQS) (17). G4-seq revealed many noncanonical G4 such as exceptionally long loop with >7 bases or bulges in G-tracts. G4-seq was conducted later for 12 model species and revealed that the enrichment of G4 in promoters is unique to mammals including human and mouse but mostly absent in other organisms including *Drosophila* (18).

A technique bearing conceptual similarity, known as rG4-seq, was subsequently devised for the study of RNA G4s across the entire transcriptome (19). The secondary structure present on template RNA obstructs continuous DNA synthesis carried out by reverse transcriptase. By

evaluating the stop sites of reverse transcription under conditions that either favor or disfavor G4 formation, rG4-seq successfully pinpointed thousands of potential sites for RNA G4.

The development of G4 structure-specific antibody (BG4) enabled probing the presence of native DNA G4 structures in cells (20). BG4-based ChIP-seq (G4 ChIP-seq) revealed landscape of G4s in various cell lines (21). Different cell types exhibit a substantial number of distinct G4 sites, suggesting that DNA G4 structures are cell type-specific and thus cell state-specific (22).

There has been considerable skepticism about the existence and biological functions of G4, especially RNA G4, in human cells (23). In order to comprehensively understand and investigate the presence of G4 structures within cells, a variety of visualization methods have been developed.

One of these methods involves the use of BG4 antibody that binds to G4 with nanomolar affinity. It has been employed to visualize and quantify both DNA G4 and RNA G4 in various states of human cells (20,24). Notably, the distinct locations of DNA G4 on chromosomes and the observed frequency changes following treatment with G4-interacting ligands suggest the existence and biological significance of DNA G4. Similarly using BG4, RNA G4 has been confirmed to be primarily located in the cytoplasm of human cells and can be stabilized by an RNA-specific G4 ligand.

In addition, small-molecule fluorescent dyes, such as thiazole orange and thioflavin T (ThT), have been developed to target G4 structures (25). Thiazole orange can achieve a high fluorescence signal upon binding to G4, but it exhibits low selectivity compared to other nucleic acid structures (26). On the other hand, ThT, originally reported for the identification of amyloid fibrils, demonstrates much higher fluorescence selectivity upon binding to G4 and has been used to visualize both DNA and RNA G4 structures in the cellular context.

1.1.3. Small-molecule G4 ligands

The mapping of G4 in DNA and RNA has revealed an enrichment of this nucleic acid secondary structure in key regulatory regions, suggesting a significant role for G4 in cellular processes. Research has shown that G4 structures are closely associated with chromosomal homeostasis, genome maintenance, proto-oncogene regulation, and the expression of cancer-related proteins (27). This correlation positions G4 as a promising therapeutic target.

G4 ligands are small molecules that either stabilize or promote the formation of G4 structure. The first therapy-oriented G4 ligand, 2,6-diamidoanthraquinone, was reported in 1997 that inhibits human telomerase through targeting $(TTAGGG)_n$ telomeric G4 (28). Subsequent studies have extended to identify small molecules targeting gene promoters (8,29). *In vitro* experiments showed transcription inhibitory effect of small molecule TmPyP4 that binds to G4 on c*MYC* oncogene possibly via G4 structure located in its promoter region.

Small-molecule G4 ligands typically contain an aromatic structure, which promotes π-π stacking interactions with the G-quartet (30). The X-ray crystal structure of the small molecule daunomycin, bound with the G4 folded by four strands of d(TGGGGT), showed the stacking of ligands on the terminal G-quartet (31). In-solution NMR analysis of pyridostatin (PDS) and its derivatives, complexed with the G4, revealed that the interaction is not limited to the aromatic rings with the G-quartet through π-π stacking (32). The aliphatic amine side chains also interact with the phosphate backbone, thereby further enhancing the affinity of PDS with the G4 structure.

The potential of G4 in the prevention of diseases, particularly cancer, has been a major driving force behind the discovery of small-molecule G4 ligands. To date, over 3700 small-molecule ligands have been cataloged in the G-quadruplex Ligands Database (https://www.g4ldb.com/) (33). With the ongoing development of ligand-screening methods, G4-

targetting small molecules will continue to serve as powerful tools for understanding the molecular mechanisms of G4 in biological processes and provide a novel angle in drug discovery.

1.2. G4 in transcriptional regulation

1.2.1. In silico and in vivo analysis showed enrichment of G4 within transcription regulatory elements

PQS identified using the consensus motif $G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}N_{1-7}$ were found to be highly enriched in the promoter regions of human genes when compared to the rest of the genome (16). Furthermore, more than 40% of human genes were found to contain one or more quadruplex motif. Additional computational analysis uncovered that these potential G4-forming sequences are not only limited to regions proximal to the transcription start site (TSS), but also found in other functional elements such as enhancers, conserved transcription factor-binding sites and nuclease-hypersensitive sites. Another line of evidence supporting the importance of G4 in biological processes comes from an evolutionay conservation analysis. By comparing the genome of *S. cerevisiae* with other yeast genomes, Carpra *et al.* determined that G4 DNA motif is significantly more conserved than expected by chance at the motif-level (34). Moreover, when compared to neighboring nucleotides, those located within the G4 motif exhibit a higher average conservation score, with a statistical significance of $p < 2.2x10^{-16}$. This strong correlation between the DNA secondary structural motif and regulatory elements from *in silico* analysis alludes a regulatory function of DNA G4 in gene transcription.

The *in vivo* confirmation of DNA G4 formation and its correlation with transcription were achieved following the development of two G4-specifc single-chain variable fragment antibodies, HF2 and BG4. Chromatin immunoprecipitation using HF2 antibody followed by sequencing in MCF7 cells identified 768 reproducible peaks, the majority of which possess G4-

forming potential (35). Motif analysis of the most enriched peak sets revealed consensus sequence of $G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}N_{1-7}$, and two example sequences were confirmed through CD spectroscopy analysis. To explore the regulatory function of promoter G4, a set of 10 genes were selected, based on HF2 enrichment within 1 kb of TSS regions. The G4-binding small-molecule ligand PDS was applied to MCF7 cells to assess changes in the expression of these selected genes. Significant altercations in gene expression were observed for 6 out of 8 G4-containing genes while two control genes without HF2 enrichment showed no change. Interestingly the alterations in gene expression were not unidirectional with both up- and down-regulation being observed following PDS treatment. This pattern suggests that G4 may play a dual role in gene expression, acting as both activator and repressor.

BG4, an antibody developed later, exhibited nanomolar affinity for G4 structures and high selectivity towards single-stranded DNA and RNA (20). BG4 ChIP-seq revealed more than 10,000 high-confidence G4 peaks in the human epidermal keratinocyte HaCaT cell line (22). In addition to the canonical G4 motifs, BG4 ChIP-seq also identified G4 tracts with longer loops or bulges, revealing diversity in cellular DNA G4 sequences. Together with two nucleosome-depleted region mapping assays, formaldehyde-assisted isolation of regulatory elements with sequencing (FAIRE-seq) and the assay for transposase-accessible chromatin with sequencing (ATAC-seq), BG4 ChIP-seq revealed that ~98% of *in vivo* G4s are predominantly located in those regions. This enrichment is not merely due to open chromatin accessibility for BG4 binding, as confirmed by further analysis.

Moreover, those genes with depleted nucleosome and G4-forming sequences still exhibited significant differences in expression regarding to the presence of G4 ChIP-seq in promoter regions. The observation that G4 folding state, apart from chromatin accessibility and

GC enrichment, affects transcription levels further substantiates the regulatory role of DNA G4 in the gene transcription.

1.2.2. Functions of DNA G4 in transcriptional regulation

DNA G4 structures are well-documented "roadblock" to DNA replication, as shown by primer extension experiments and polymerase stalling assays conducted *in vitro* (11,36,37). *In vivo* study, involving helicase deficiencies, has also demonstrated the impediments posed by G4 structures to replication. Deletion of FANCJ, a helicase that can recognize and unwind G4 structures, in *Caenorhabditis elegans*, resulted in accumulation of small deletions located upstream of potential G4 sequences (38). Similarly, DNA transcription mediated by RNA polymerases has also been hypothesized to be hindered by G4 structures. This notion is supported by earlier *in vitro* DNA transcription experiments which revealed RNA synthesis blockages when a G4 motif was present in the transcribed region (39).

RNA polymerase II-mediated transcription proceeds through three stages: initiation, elongation, and termination (40). An additional checkpoint, "promoter proximal pausing" occurs between initiation and elongation to regulate transcription. Analysis of RNA polymerase II ChIP-seq profiles showed a correlation between G4 motif flanking TSS and this pausing (41). This correlation supports the idea that G4 structures can act as transcription repressors by obstructing essential transcription machinery.

However, numerous G4 structures are associated with actively transcribed genes, the above model doesn't elucidate the mechanism by which G4s serve as transcription enhancers (22). One plausible explanation is that the formation of G4 structures can stabilize R-loops. As duplex DNA unwinds and RNA polymerase advances along the template strand, a three-stranded nucleic acid configuration emerges (42). The RNA:DNA hybrid, known as R-loop, was generated and found to be most stable if formed in the C-rich regions on the DNA strand (43). This is

compatible with G4 formation in guanine-rich sequences. Placing a G4-forming sequence on the non-template strand to facilitate R-loop formation has been observed to significantly boost transcription by RNA polymerase with increasing RNA production and enhanced RNA polymerase elongation (44). However, this effect was reversed when the G4 motif was positioned on the template strand of DNA, further supporting the hypothesis that G4 formation can augment gene transcription through stabilizing R-loops.

1.2.3. DNA G4-interacting proteins that correlate with transcription activation

An alternative model proposes that G4 DNA acts as docking site for nuclear proteins, particularly transcription factors. A variety of proteins have been identified to interact with DNA G4 structures, with human specificity protein (SP1) being one of the most extensively studied. SP1, a basal transcription factor, is ubiquitously expressed in mammalian cells and regulates numerous housekeeping genes (45). It contains three tandem C2H2 zinc finger domains, which are responsible for DNA binding, at the C-terminus. The SP1's consensus binding motif was originally recognized as 5'-GGGCGG-3', termed GC-box. Yet, a subsequent ChIP assay combined with microarray revealed only a minority of the SP1 binding sites contained the consensus motif, hinting alternative binding patterns (46). Sequence analysis of *c-KIT* promoter, one of SP1 binding sites, possesses a G-quadruplex forming sequence and later was confirmed using CD spectroscopy. Biochemical assays further revealed SP1's higher affinity for the G4 structure in the *c-KIT* promoter relative to its ssDNA counterpart. Additional research illustrated SP1's interaction with the *HRAS* promoter, which also contains a G4-forming sequence (47). Intriguingly, obstructing G4 formation through point mutations led to a five-fold increase in transcription, suggesting that the HRAS G4 might function as a transcription repressor. Another transcription factor, the Myc-associated zinc-finger protein (MAZ), was also found to bind G4 structures. MAZ has been linked with the KRAS gene promoter, which houses a G4 forming

sequence (48). In this context, site-specific mutations that abrogate G4 formation result in KRAS down-regulation, whereas stabilizing G4 using guanidine-modified phthalocyanines elevates transcription activity. The contrasting outcome of G4 stabilization and disruption experiments underscore the versatile role G4 structures play in transcription regulation.

Another set of G4-interacting proteins may modulate transcription activity by modifying chromatin environment. The Bromo domain containing protein 3 (BRD3), identified as one of the top hits in microarray screening of the G4-interactome, was later validated to engage with G4 structures both *in vitro* and *in vivo* (49). BRD3 is a well-known chromatin reader that binds to acetylated histones and aids in the recruitment of RNA polymerase (50). The pronounced co-localization between BRD3 and BG4 ChIP-seq suggests a potential role for G4 in the chromatin remodeling processes orchestrated by BRD3 (51).

Apart from histone modification, G4 structures have also been linked with DNA methylation, a widely researched epigenetic mark (52). Methylation on the C5 position of cytosine, especially near the transcription start sites, plays a pivotal role in gene expression. Through a combination of BG4 ChIP-seq and whole-genome bisulfite sequencing, a technique for measuring methylation levels across the genome, Mao et al. discovered that endogeous G4 structures are closely linked with hypomethylation at CpG islands (53). Further overlapping analysis indicated an enrichment of DNA methyltransferase 1 (DNMT1) at these G4 sites, a finding supported by subsequent biochemical assays. The counterintuitive observation that an enzyme responsible for DNA methylation is located in hypomethylated regions led to the hypothesis that the binding of DNMT1 by G4 structure might inhibit its enzymatic activity.

The multifaceted roles of G4 structures in transcription regulation highlight the need for methods to identify G4-interacting proteins. Several approaches have emerged, including proteomics pulldown, computation analysis and genetic screening.

Using carefully designed biotin probes, cellular proteins are incubated and subsequently captured with streptavidin beads based on their affinity toward either G4 sequences or control sequences. The eluted proteins are then subjected to liquid chromatography with tandem mass spectrometry analysis (LC-MS/MS), enabling the detection of potential G4-binding proteins. With advances in analytical proteomics methods, a variety of quantitative strategies have been adopted. Label-free quantification, notable for its cost-effectiveness and absence of tagging, has led to the identification of G4-binding proteins like nucleolin and PARP (54,55). The stable isotope labeling by amino acid in cell culture (SILAC)-based quantitative proteomics, on the other hand, offers a more accurate measurement of differences in protein affinity to either G4 or control sequences. This method has facilitated the discovery of a number of G4-interacting proteins, including SLIRP, YY1 and VEZF1 (56).

Computational analysis of protein binding sites to determine the enrichment of G4 motifs within these binding sites also yields putative G4-binding proteins. Results from SP1 ChIP-on-chip analysis revealed an 87% overlap with putative G4 sequences and exhibits selective binding towards G4 region in the *c-KIT* promoter (46). Overlapping analyses of XPB, XPD, and DNMT1 highlighted an enrichment of either G4 motifs or endogenous G4 structures at their respective binding sites (57). These insights guided the biochemical assays to confirm their physical interactions. Through a comprehensive analysis combining ChIP-seq datasets from human chronic myelogenous leukemia cells (K562) and hepatocellular carcinoma cells (HepG2) with BG4 ChIP-seq, researchers identified a preference for localization of transcription factors within endogenous G4 structures compared to double-strand DNA (58). This suggests that G4 structures might serve as a protein docking hub for transcription.

Recognizing the crucial roles that G4 and its interaction proteins play in cellular processes, genetic screenings employing shRNA and small-molecule G4 ligands have been used

to pinpoint proteins involved in G4-associated biological pathways (59). Using an array of 132,000 shRNAs targeting the human proteome and two distinct G4 ligands, PDS and PhenDC3, Zyner et al. identified 758 G4-sensitiser genes (59). These genes are prominently enriched in five KEGG pathways: cell cycle, ribosome, splicesome, ubiquitin-mediated proteolysis and DNA replication. Furthermore, they discovered ATP-dependent RNA helicase DDX42 as a novel G4 binding protein.

1.2.4. DNA G4 with long-range DNA interactions

Mapping of endogenous G4 reveals the presence of these secondary structures not only within the promoters of actively transcribed genes but also within introns and intergenic regions. This suggests G4's potential distal regulatory role in transcription, raising questions about its role in 3D genome organization.

Human chromatin's intricate organization within the nucleus ensures proper gene expression and accurate genome replication (60,61). At the highest level, the genetic material is segregated into A/B compartments, distinguishing actively transcribed euchromatin from repressed heterochromatic regions. Beneath that, topologically associated domains (TADs) demarcate locally interacting regions, with infrequent interactions between neighboring TADs. Chromatin looping represents the most granular level of 3D genome architecture, containing crucial long-range DNA interactions in epigenetic transcription control, i.e., promoter-enhancer contacts (62). Enhancers are sequences that bolster transcription by recruiting auxiliary proteins. Unlike promoters situated immediately upstream of transcription start site, enhancers can be distant in the genome sequence from target gene. The Shh promoter-enhancer contact serves as a prototypic example: located roughly 850 kb downstream of Shh promoter, Shh enhancer modulates Shh expression during mouse limb development (63). Its deletion caused a dramatic reduction in Shh expression, leaving the expression of four nearby genes unaffected (64).

Several proteins, including CCCTC binding factor (CTCF), Yin Yang 1 (YY1) and cohesin, and the mediator complex, have been linked to chromatin interactions (65). The "loop extrusion" model postulates genome division into regulatory domains facilitated by CTCF, cohesin, and other cofactors (66). The cohesin complex, comprising SMC1, SMC3, RAD21 and STAG, forms a ring-like structure, entrapping DNA strings. CTCF, found to colocalize with cohesin, acts as a brake preventing the cohesin complex from sliding along DNA fibers. Both CTCF and cohesin are enriched at TAD boundaries, indicating their insulating role in preventing inter-domain interactions (67).

However, integrative analysis of Hi-C and BG4 ChIP-seq reveals that TAD boundaries demarcated by CTCF not only possess the CTCF consensus motif but also G4 structures (68). Binding profiles of proteins involved in cohesin complex from ChIP-seq indicated greater recruitment of RAD21 and SMC3 to G4-containing boundaries compared to non-G4 counterparts. Moreover, interactions between adjacent TAD boundaries are more frequent when G4 structures are present. These findings suggest that G4, as DNA secondary structures, might enhance the CTCF-mediated loop intrusion and influence 3D genome organization.

More intriguing evidence that G4 is involved in DNA looping comes from the work exploring the interaction between G4 and YY1, shown to facilitate promoter-enhancer loops (69). Lin et al. first verified direct YY1-G4 interaction using the electrophoretic mobility-shift assay and florescence anisotropy. Subsequent ChIP-seq experiments demonstrated a high co-localization of YY1 with endogenous G4 structures. G4 ligands treatment, using PDS and TMPyP4, significantly reduced YY1 binding in G4 regions, corroborating *in vivo* interaction. Moreover, these treatments also disrupted YY1-mediated long-range DNA interactions, as determined via HiChIP-seq. Site-specific mutation using CRISPR-Cas9 on selected regions further proved that the disruption of G4 folding perturbed the YY1 binding and looping. In a

detailed studied example, the TRMT12 promoter connects to a remote G4-containing region. Small-molecule G4 ligand treatment disrupted not only YY1 binding at the distal G4, but also the YY1-mediated promoter-enhancer loop. This diminution led to a significant decrease in TRMT12 mRNA expression.

While extensive studies have been conducted for G4's role in transcription factor binding and epigenetic regulation, its function in long-range DNA interactions remains under-explored. However, these regulatory pathways likely operate synergistically rather than isolated. With the inspiring Hi-C analysis and the enlightening YY1 example, higher-order chromatin structures may offer additional insights into how DNA G4s act as transcription regulator.

1.3. RNA G-quadruplex

1.3.1. Skepticisms of RNA G4 formation within cells

Analogous to the DNA G4 structure mapping, RNA G4 was initially identified using computational analysis with the consensus sequence motif $G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}N_{1-7}$ (16,70). Potential RNA G4 sequences were found to predominantly appear in 5' and 3'-UTRs, regions crucial for post-transcriptional regulation. The above mentioned rG4-seq probed the transcriptome-wide localization of RNA G4s by reverse transcriptase stalling in environments favoring G4 folding (19). This *in vitro* mapping revealed the prevalence of RNA G4 in the human transcriptome. However, subsequent studies that probed RNA folding states in cells using dimethyl sulfate or SHAPE (selective 2'-hydroxyl acylation analyzed by primer extensions) reagent combined with next-generation sequencing, produced contrasting findings (71). This steady-state sequencing method revealed that, while the physiological metal cation environment should favor its formation, RNA G4s appeared to be predominantly unfolded in human cells.

These discrepancies between *in vitro* and *in cellulo* results prompt the hypothesis that RNA G4s form transiently and dynamically. The RNA G4-specifc fluorescent probe, QUMA-1, was employed to observe the folding and unfolding processes (72). Its high cell permeability and low cytotoxicity enabled real-time visualization of RNA G4 in live cells. Observations of highly dynamic events including folding/unfolding in a few seconds and merging/splitting of QUMA-1 foci, corroborating the theory of transient RNA G4 formation. Further evidence supporting the presence of RNA G4 *in vivo* came from transcriptome-wide mapping of transient RNA G4 in cells. This method, termed G4RP-seq, captured RNA G4 with the G4-specific small-molecule ligand, BioTASQ (73). After crosslinking to snapshot RNA in either its folded or unfolded state, the biotinylated ligand enabled the enrichment of G4-containing RNA. G4RP-seq not only revealed a wide-spread distribution of G4 in protein-coding RNAs but also demonstrated a conspicuous absence of G4 in long non-coding RNAs. These findings strengthen the argument of transient RNA G4 formation and hint at a potential protein machinery regulating this folding/unfolding process.

The first protein identified for resolving RNA G4 structures is DEAD/H box protein 36 (DHX36) (74). Initially discovered as a resolvase for DNA G4, DHX36 was later confirmed to possess the ability to unwind RNA G4 (75). Numerous experiments demonstrated its involvement in mRNA and long-non-coding RNA processing by targeting and resolving RNA G4 structure. Detailed fluorescence resonance energy transfer (FRET) experiments revealed that DHX36 binding induces repeated cycles of ATP-independent unfolding and ATP-dependent refolding before it is detached from RNA G4 (76,77). This unique unfolding/refolding mechanism of DHX36 in resolving RNA G4 substantiated the notion that RNA G4 structures are transient and highly dynamic in cells.

1.3.2. Function of RNA G4 in RNA processing

The prevalence of RNA G4 structures in regulatory regions, notably the 5'- and 3'-UTRs of pre-mRNAs and mRNAs, underscores their significance in RNA metabolism and translation. Below, we will discuss the diverse functions of G4 in RNA biology, and how RNA-binding proteins serve as determinants in these processes.

Transcription termination is intricately tied to the enzymatic cleavage of nascent RNA, subsequently followed by the attachment of poly(A) tails (78). This polyadenylation of newly synthesized transcripts directly impacts the stability and proper maturation of RNA molecules. In eukaryotic cells, the installation of poly(A) at the 3' end is orchestrated by four major protein complexes: cleavage polyadenylation specificity factor (CPSF), cleavage stimulation factor (CstF), and cleavage factors I and II (79). The central motif AAUAAA marks the recognition site of CPSF and guides RNA cleavage (80). For those transcripts bearing G4 sequences at the poly(A) signal regions, e.g., *TP53* mRNA, heterogeneous nuclear ribonucleoprotein H/F (hnRNP H/F) regulates the polyadenylation together with the essential protein complexes (81,82). In the example of *TP53*, 3'-end processing is diminished in mRNA compared to transcripts devoid of G4s under optimal physiological conditions (83). However, in the face of DNA damage or genotoxic stress, canonical polyadenylation is inhibited by the entrapment of Cstf complex within a repressed protein complex. By contrast, hnRNPH/F's affinity for G4 facilitates the recruitment of CstF and enables efficient 3'-end processing of *TP53* pre-mRNA. This consequent up-regulation of *TP53* stimulates the expression of proteins geared towards stress-response and damage control.

Function of RNA G4s in alternative splicing have also been demonstrated. Alternative splicing serves as an essential tool to increase transcriptome complexity and plays important roles in cell differentiation, organ development and human diseases (84). The Fragile X mental

retardation protein (FMRP) has been identified to interact with RNA G4 using *in vitro* binding assays (85). An association between FMRP's binding site on its pre-mRNA (FMR1) and the relative expression of FMRP isoforms was revealed, where amplified FMRP bindings are correlated with the longer isoform's heightened expression (86). Sequence analysis unveiled two G4 motifs within this region. Notably, mutations abolishing G4 forming potential curtailed the splicing enhancer activity of FMRP, skewing the relative proportions of FMRP isoforms. Another example is the G4 in human telomerase reverse transcriptase (hTERT), which has been shown to act as a splicing suppressor (87). Stabilizing RNA G4 with G4-specific ligand hindered hTERT splicing. Coupled with further genome-wide analysis highlighting G4's enrichment proximal to splice junctions, these investigations substantiated the role of RNA G4 in modulating mRNA synthesis.

1.3.3. Functions of RNA G4 in translation

mRNA translation represents the final phase of the central dogma, wherein genetic information is transformed into a protein sequence. The molecular mechanisms underlying mRNA translation in mammalian cells are intricate and still undergoing active research. Translation initiation commences with the mRNA cap being recognized by the eukaryotic translation initiation factor 4E (eIF4E) complex (88). Following this, the eIF3 translation initiation complex unites with the 40S ribosomal subunit, giving rise to the preinitiation complex 43S. As the eIF2-GTP-$tRNA_i^{Met}$ complex integrates with 43S, it scans the 5'UTR for the AUG start codon. Once the start codon is identified, the full-fledged translation machinery assembles, advancing to the elongation phase of translation.

Thousands of RNA G4 forming sequences have been identified within the 5'-UTR of the human transcriptome. Detailed examinations of numerous human mRNAs, harboring G4 forming-sequences in their 5'-UTR, underscored the repressing role of G4 in translation (89). For

instance, the RNA G4 structure in NRAS proto-oncogene mRNA has been shown to obstruct the scanning process of initiation in *in vitro* translation assays. Moreover, this inhibitory impact is intrinsically linked to the stability and location of the G4 sequence on the 5'-UTR, as validated both *in vitro* and in cells. Observation in various mRNAs, including ZIC-1, TRF2 and MT3-MMP, raised the proposition that RNA G4s within the 5'-UTR act as barriers during the scanning process of translation initiation (90).

In contrast, instances where RNA G4 in the mRNA 5'-UTRs enhance translation have also been reported, including pivotal genes associated with cancer progression such as vascular endothelial growth factor (VEGF) and NF-E2-related factor 2 (NRF2). Sequence analysis of human VEGF's 5'-UTR revealed a G4 structure nestled within the internal ribosome entry site (IRES), which allows cap-independent translation initiation (91). RNase footprinting and CD spectroscopy validated the G4 folding *in vitro*. A dual-luciferase assay, employing plasmids with or without this G4 sequence in VEGF 5'-UTR, demonstrated an essential role of this G4 structure in VEGF translation. Translation initiation of NRF2 was shown to be induced by the recruitment of 18S rRNA to the IRES within the 5'-UTR (92). Subsequent research highlighted an RNA G4 structure proximal to the NRF2's IRES and found the crucial role of this RNA G4 structure in NRF2 expression , which is essential in antioxidation and detoxification, under oxidative stress (93). The close correlation between IRES and RNA G4 suggests significance of RNA G4 in cap-independent translation.

While most research on RNA G4 in translational regulation highlighted G4 structures at 5'-UTR, there are a few studies delving into the roles of RNA G4s within 3'-UTRs and coding regions in modulating translation. An identified G4 structure within the 3'-UTR of proto-oncogene PIM1 acts as a translation repressor (94). Meanwhile, destabilizing the G4 structure in the transcribed region of Epstein-Barr virus-encoded nuclear antigen 1(EBNA1) augments mRNA

translation of EBNA1, with small-molecule G4 ligand that stabilizes G4 structure dampening EBNA1 protein synthesis (95).

Collectively, these studies revealed multifaceted roles of RNA G4 structure in mRNA translation. While the majority of studies posit the G4 structure as an impediment to translation, a subset showcases its alternative role as essential elements in the production of selected proteins. With advances in methodologies for RNA biology and the further identification of crucial RBPs, the intricate mechanisms by which RNA G4 regulates translation can be further elucidated.

1.4. G-quadruplex in telomere maintenance

1.4.1. Telomere and Telomeric G4

Telomeres are conserved DNA-protein complex capping the ends of chromosomes with tandem guanine-rich repeats (96). A six-member protein shelterin complex protects telomeric DNA from inappropriate processing by DNA repair pathways (97). The length and composition of telomeres vary among organisms. Human telomeres, for instance, consist of TTAGGG repeats and typically span between 10 to 15 kb (98). Beyond the double-stranded region of several kilobases, telomeres also possess single-stranded G-rich sequences at termini.

Telomeres are well-known for their correlation with aging (99). The gradual shortening of telomere length occurs during the division of somatic cells due to end replication problem (100). During the DNA replication of the lagging strand, multiple RNA primers function as starting points for polymerase engagement and are later degraded post DNA synthesis. The residual gap at the terminal end remains unfilled, resulting in the loss of a short DNA segment. This shortening is postulated to act as the first replicative defense against tumour formation (101).

In early human development, stem cells preserve telomere length through the ribonuclear protein known as telomerase (102,103). The telomerase RNA component (TERC) serves as a

template for restoring telomeric repeats, with telomerase reverse transcriptase (TERT) conducting the DNA synthesis (104). This process is supported by accessory proteins such as dyskerin, GAR1, NHP2, NOP10 and TCAB1. Telomerase is up-regulated in 85-90% of human cancers and plays essential roles in cancer progression.

On the other hand, in ~10-15% of tumors, telomeres are elongated via an alternative mechanism termed alternative lengthening of telomeres (ALT) (105). ALT was observed from telomere length heterogeneity displaying extremely long (> 50kb) and short (< 5kb) telomeres (106). Sequencing of ALT-positive cancer genomes has revealed prevalent mutations in the α-thalassemia/mental retardation syndrome X-linked protein (ATRX) and the death domain-associated protein (DAXX) (107). Both are responsible for the deposition of histone variant H3.3 within telomeric regions (108). The dysregulated chromatin state due to aberrant ATRX-DAXX mutations up-regulates TERRA level (109). This, in turn, stimulates the telomeric recombination from a non-homologous chromosome (110). The detailed mechanism of ALT in human cells are under active investigation, with G4 structures emerging as potentially important players.

The telomeric DNA G4 was the first identified biologically relevant G4 structure using single-stranded oligonucleotides representing the telomere sequences of ciliated protozoa (111). A comprehensive analysis of telomeric sequences across 15 species showcased *in vitro* G4 formation capabilities in nearly all of them (112). Autoradiography of human cell metaphases treated with radio-labeled G4 ligands demonstrated G4 formation at the ends of metaphase chromosomes (113). The G4-specific antibody BG4 enabled *in vivo* visualization of G4 structures and unveiled telomere G4 formation togethering with fluorescence in situ hybridization targeting telomeric DNA (20).

1.4.2. Function of DNA G4 in telomere

The conservation of G-rich sequence in telomeres hints at the selective evolutionary advantage of G4 formation in telomeric regions, possibly playing a pivotal role in telomere biology. However, due to the limit of resolution, the precise locations of G4 structure within telomeric regions remain largely undermined. Possible G4 structures could form in the single-stranded overhang, during DNA replication of the telomere or during the transcription of TERRA.

G4 formation in single-stranded overhangs has been speculated as a protective cap against nucleases. A seminal study conducted in *S. cerevisiae,* involving a mutation in the telomere capping protein Cdc13, showed that enhancing G4 formation through various means, such as G4-stabilizing ligands, expression of G4-stabilizing proteins and deletion of G4-resolving protein, mitigated the growth defect of the mutant (114). Subsequent experiments employing the G4-specific antibody BG4 to measure G4 structure revealed an increase in G4 signal at the telomeres in Cdc13 mutant yeast (115). These research supports the idea that G4 can serve as alternative caps to safeguard telomere DNA.

Similar to studies of G4 formation in DNA replication, G4 might hinder the replication of telomeric regions. Treating cells with small-molecule G4-stabilizing ligands, such as PDS, leads to replication stalling at telomeres of human cells (116). However, small molecule analysis of replicated DNA (SMARD) indicated comparable replication fork progression between telomeric and non-telomeric regions (117). This finding suggests the involvement of proteins that resolve G4 structures during replication. Several studies have identified helicases, including BLM, WRN and FANCJ, that can unwind G4 structure (118). The absence of these proteins manifests in a “fragile telomere” phenotype with dysfunction in telomere replication. An elevated presence of BG4 loci at telomeres in cells with BLM or WRN defects further substantiated the significance of G4 unwinding during telomere replication (119,120).

Despite the heterochromatin states within telomere regions, TERRA was transcribed in a regulated manner by RNA polymerase II from the C-rich strand (121). Its G-rich characteristic allows TERRA to form G4 structures both *in vitro* and in cells. Studies have shown its potential role in several cellular processes and diseases. The translocated in liposarcoma (TLS) protein has been identified to bind telomeric DNA G4 and TERRA G4 structure both *in vitro* and *in vivo*, through its C-terminal Arg-Gly-Gly (RGG) domain (122). Intriguingly, overexpression of the full length or C-terminal domain of TLS leads to telomere shortening, implying a possible mechanism of G4-mediated telomere length regulation by TLS. Further experiments showed a progressive increase in histone methylation, including H3K9me3 and H4K20me3, after TLS overexpression (123). Such findings suggest that G4 structures, whether in telomeric region or TERRA, serve as scaffolds for TLS binding to regulate telomere length and chromatin state.

1.5. Next-generation sequencing techniques in molecular biology studies

Sanger sequencing, developed in the 1970s, marked a pivotal shift in biology, transforming the field into the era of genetics and genomics (124,125). Utilizing chain termination and fragmentation methods, Sanger sequencing facilitated the completion of the first-ever human genome sequence in 2004 (126). With an increasing demand for greater sequencing throughput and cost reduction, next-generation sequencing (NGS) was proposed to replace the traditional Sanger methods that could run millions of sequencing reactions simultaneously and eliminate the need for electrophoresis.

The first NGS technology was released in 2005 by 454 Life Sciences, capable of generating approximately 20 million reads, each 110 base-pairs long (127). A year later, the Solexa/Illumina sequencing platform was introduced, eventually becoming the most widely utilized NGS technique. Today, the Illumina NovaSeq X system can sequence up to 20 billion

reads, each 150 base-pairs long, in one flow-cell, enabling the sequencing of 24 complete human genomes.

With continuous advancements in sequencing instruments and decreasing costs, a variety of NGS-based methods have been developed to tackle biological questions across various fields, including genomics, epigenetics, transcriptomics, and epitranscriptomics. In the following section, we will briefly introduce several techniques that are ubiquitously applied to study the functions of G4 in cells that can provide profound insights into their functions and potential as therapeutic targets.

1.5.1. Chromatin immunoprecipitation followed by sequencing (ChIP-seq)

Numerous biological processes such as gene transcription, DNA replication, chromosome remodeling, and epigenetic regulation are intricately linked to the interactions between cellular proteins and DNA. Discovering the binding sequences of transcription factors is pivotal in understanding transcription regulation and can reveal potential therapeutic targets. Additionally, the detailed mapping of various histone modifications helps to elucidate the complex epigenetic landscape of chromatin. ChIP-seq offers a direct means to measure the genomic locations of proteins, serving as a vital tool in the exploration of these critical cellular functions (128).

The experiment procedure of ChIP-seq starts from fixing cells with formaldehyde that crosslinks protein with its bound DNA *in vivo* (128). The cells are then lysed and DNA sonicated into shorter pieces. The fragmented chromatin is immunoprecipitated with a specific antibody that captures the target protein. Finally, the protein-DNA crosslinks are reversed and the immunoprecipitated DNA is purified and analyzed using NGS method. Data analysis of ChIP-seq usually composed of quality control, reads alignment, peak calling, and downstream processing. Depending on the protein of interest, different types of downstream analysis can be performed. This may include motif finding to uncover binding patterns, genome annotation to identify

functional elements, and chromatin state determination to describe the epigenetic environment of the chromatin.

ChIP-seq has emerged as a common tool in G-quadruplex studies. As previously mentioned, the localization of G4 within the cellular context has been achieved through ChIP-seq analysis with the use of G4-specific antibody BG4 (G4 ChIP-seq) (22). ChIP-seq experiments of XPB and XPD, two essential helicases involved in nucleotide excision repair showed enrichment at G4 motifs that associate with downstream signaling pathways (57). By examining the overlapping percentage of the binding site of DNA-binding proteins with *in vivo* mapping of G4, researchers were able to identify DNA methyltransferase 1 (DNMT1) as a G4-binding protein (53). This led to the proposal of a G4 sequestration model to explain the observed low DNA methylation levels at G4 regions.

As a robust and adaptable technique, ChIP-seq can provide valuable insights into the function of G4 in the complex gene regulatory networks, which opens up new avenues of research in genomics and epigenetics.

1.5.2. Crosslinking and immunoprecipitation following by sequencing (CLIP-seq)

RNA is seldom found in a naked form, as proteins start to interact with RNA as early as transcription initiation. Ribonucleoprotein (RNP) complexes play crucial roles throughout the entire life cycle of RNA, including processes such as transcription, splicing, cytoplasmic export, translation and degradation (129). Dysfunctions in RNA binding proteins are associated with various diseases including neurodegeneration, auto-immune disorder and cancer. To deeply understand the biological roles of RNA-binding proteins, it is essential to identify the types and sequences of RNAs to which they bind within cells, particularly in different environmental and developmental states.

Similar to ChIP-seq, CLIP-seq is used to characterize the interaction between specific protein and RNA (130). However, CLIP-seq employs ultraviolet (UV) light instead of formaldehyde to irreversibly crosslink proteins to their bound RNA. General procedures of CLIP-seq include preparation of crosslinked cell lysate, RNA fragmentation, immunoprecipitation of specific protein cross-linked RNA, fragment purification and cDNA synthesis. Several variations of CLIP-seq have been developed including photoactivatable ribonucleoside-enhancer CLIP (PAR-CLIP), individual-nucleotide resolution CLIP (iCLIP) and enhanced CLIP (eCLIP) (131,132). PAR-CLIP identify crosslinking sites with nucleotide resolution through the analysis of mutations induced by crosslinked ribonucleosides, while iCLIP and eCLIP determine the binding sites by capturing the exact point where reverse transcription terminates due to crosslinking.

Fragile X mental retardation protein (FMRP) is highly expressed in the brain and known to be correlated with Fragile X syndrome, a genetic disorder linked to intellectual disabilities. By analyzing the CLIP-seq dataset of FMRP, researchers were able to demonstrate that the FMRP's recognition of RNA G4 is vital to the transport of G4-containing transcripts to neurites thus implicates the role of RNA G4s in neuronal function (85). Furthermore, bioinformatics analysis of hnRNPF-binding regions revealed a significant enrichment of predicted G-quadruplex sequence. This G4 bias has been found to occur near alternative exons regulated by hnRNPF (133). Subsequent experiments confirmed the role of RNA G4 in alternative splicing.

These discoveries not only underscore the importance of RNA G4 in biological processes, but also highlight the power of CLIP-seq as a tool to unravel complex RNA G4-protein interactions. The vital role of RNA-binding proteins in the formation and biological functions of RNA G4 has stimulated the discovery of RNA G4-binding proteins. As a direct and informative

technique, CLIP-seq can be used in conjunction with complementary methods to offer a comprehensive view of RNA binding proteins in RNA G4 biology.

1.5.3. ChIA-PET and HiChIP-seq

The eukaryotic genome is organized in a hierarchical fashion ranging from broader A-B compartment to specific enhancer-promoter contacts (134). High-order chromatin organization plays an important role in essential biological processes and disease development. To understand 3D genome architecture, two main approaches have been developed (135). The first method leverages the use of fluorescent in situ hybridization coupled with high-resolution microscopy, allowing for the visualization of specific looping interactions within the chromatin. The second category is mainly derived from the technology known as chromosome conformation capture (3C), that detect and analyze physically proximal DNA interactions.

Standard 3C assays consist of several key steps, including crosslinking with formaldehyde to capture physical contacts between chromosomal regions, digestion by restriction enzymes, ligation to connect proximal digested fragments, reverse crosslinking, DNA purification and analysis through quantitative PCR. Derived techniques from the 3C assay can be categorized into various types based on the interaction mapping range: one vs. one (3C), one vs. many (3C-seq and 4C), many vs. many (ChIA-PET, HiChIP-seq), many vs. all (Capture-3C) and all vs. all (Hi-C). Among them, Hi-C, ChIA-PET and HiChIP-seq employ high-throughput sequencing to analyze the re-ligated DNA fragments, thereby producing comprehensive information about chromosome organization on a genome-wide scale (136,137). While Hi-C allows for the measurement of the all-possible proximity ligation products, ChIA-PET and HiChIP, which include additional immunoprecipitation steps, enable targeted DNA loops detection mediated by particular protein.

Owing to the cumbersome procedures and high genome coverage requirement with Hi-C/ChIA-PET/HiChIP, there are much fewer studies reported in using 3C-based NGS assays to explore the role of G4 in the genomic context comparing to ChIP-seq and eCLIP-seq. However, the example of YY1 showed the importance of HiChIP-seq in elucidating the interplay between nucleic acid secondary structures and DNA looping. This encouraging instance showed the indispensable role of 3C-based assays in depicting high-order chromatin structures and its importance in probing the detailed mechanism by which G4 modulates gene expression.

1.5.4. ENCODE database

Since its inception in 2003, the Encyclopedia of DNA Elements (ENCODE) consortium has been employing a wide range of assays and methods, predominantly those based on sequencing, to explore gene elements (138-140). These elements include modified histones, transcription factors, chromatin regulators, and RNA-binding proteins.

The ENCODE project has resulted in the creation of over 13,000 datasets, now available on the ENCODE portal. These encompass information on proteins from a variety of species, including humans, mice, and Drosophila. In addition to its broad accessibility, one of the key strengths of the ENCODE datasets is the adherence to standard, robust, and well-established data processing pipelines. This uniformity in processing significantly reduces the complexity of downstream analysis.

By providing a comprehensive and standardized resource for researchers, ENCODE fosters deeper understanding of genomic elements and their functions, facilitating advancements in various fields of biology, medicine, and genetics. Its collaborative and open-access nature ensures that a wide community of scientists can benefit from these valuable insights and contribute to the ongoing exploration of the complex world of DNA.

1.6. Scope of this dissertation

The pervasive presence of G4 structures within regulatory domains has spurred intensified research efforts to understand their role in modulating cellular functions. Numerous investigations affirm that both DNA and RNA G4s actively partake in pivotal biological pathways, including DNA replication, transcription, RNA metabolism, translation, and telomere maintenance. The advent of native G4 mapping allows for genome-wide appraisal of G4 structures in living cells, further amplifying inquiries into their regulatory potential. Yet, while many studies have focused on individual G4-interacting proteins or specific G4-forming sequences, an overview detailing the interplay between these secondary structures and other regulatory mechanisms still awaits thorough exploration. Utilizing bioinformatics analyses of publicly accessible datasets, we delved into the multifaceted roles of G4 structures across histone modifications, three-dimensional chromatin configurations, RNA metabolism, and telomere maintenance. Our comprehensive analysis of next-generation sequencing data presents a fresh perspective, shedding light on the intricate regulatory mechanisms mediated by G4 structures.

In chapter 2, we revealed the G4 overlapping pattern with transcription factors and constructed an interaction network of potential G4-interacting proteins. This network hints at protein clusters that might have G4-dependent roles in the regulation of essential cellular processes. Moreover, we explored the potential interplay between G4 structures and histone marks, unveiling G4 structures as an active transcription mark. Together with overlapping analysis of histone modification writer and erasers, we postulated an intricated G4-dependent regulation of histone modifications.

In chapter 3, we conducted an intuitive overlapping analysis of previously published RNAPII ChIA-PET and BG4 ChIP-seq data. We observed a strong positive correlation between RNAPII-linked DNA loops and G4 structures in chromatin. Additionally, our RNAPII HiChIP-

seq results showed that treatment of HepG2 cells with pyridostatin (PDS), a small-molecule G4-binding ligand, could diminish RNAPII-linked long-range DNA contacts, with more pronounced diminutions being observed for those contacts involving G4 structure loci. RNA-seq data revealed that PDS treatment modulates the expression of not only genes with G4 structures in their promoters, but also those with promoters being connected with distal G4 through RNAPII-linked long-range DNA interactions. Together, our data substantiate the function of DNA G4 in RNAPII-associated DNA looping and transcription regulation.

In chapter 4, we employed a bioinformatic approach based on the analysis of overlap between peaks obtained from rG4-seq analysis and those detected in > 230 eCLIP-seq datasets for RNA-binding proteins generated from the ENCODE project. We identified a large number of candidate rG4-binding proteins. We showed that one of these proteins, G3BP1, is able to bind directly to rG4 structures with high affinity and selectivity, where the binding entails its C-terminal RGG domain and is further enhanced by its RRM domain. Additionally, our seCLIP-Seq data revealed that pyridostatin, a small-molecule rG4 ligand, could displace G3BP1 from mRNA in cells, with the most pronounced effects being observed for the 3'-UTR of mRNAs. Moreover, luciferase reporter assay results showed that G3BP1 positively regulates mRNA stability through its binding with rG4 structures. Together, we identified a number of candidate rG4-binding proteins and validated that G3BP1 can bind directly with rG4 structures and regulate the stabilities of mRNAs.

In chapter 5, I proposed a novel approach for identifying putative telomere-binding proteins through ChIP-seq data analysis. In light of previous identification of zinc finger (ZNF) domain proteins as telomere-binding proteins, I conducted a comprehensive enrichment analysis focused on ZNF and ZBTB proteins. Three prominent targets, ZNF24, ZNF316 and ZBTB33 demonstrated significant enrichment at telomeric regions. A detailed examination of ZBTB33

suggested a potential G4-dependent telomere binding activity. Our analysis provides a new angle on understanding proteins involved in telomere biology.

In summary, this dissertation introduces a pioneering bioinformatics approach to investigate the intricate interplay between G4 structures and other cellular regulatory mechanisms. This innovative methodology facilitates connections between DNA secondary structures, vital epigenetic marks, and long-range DNA interactions. Additionally, it has been instrumental in identifying new RNA binding proteins and probing potential G4-dependent telomere-binding proteins. Collectively, these insights underscore the regulatory significance of G4 structures and shed light on their intricate roles in the regulation of cellular processes.

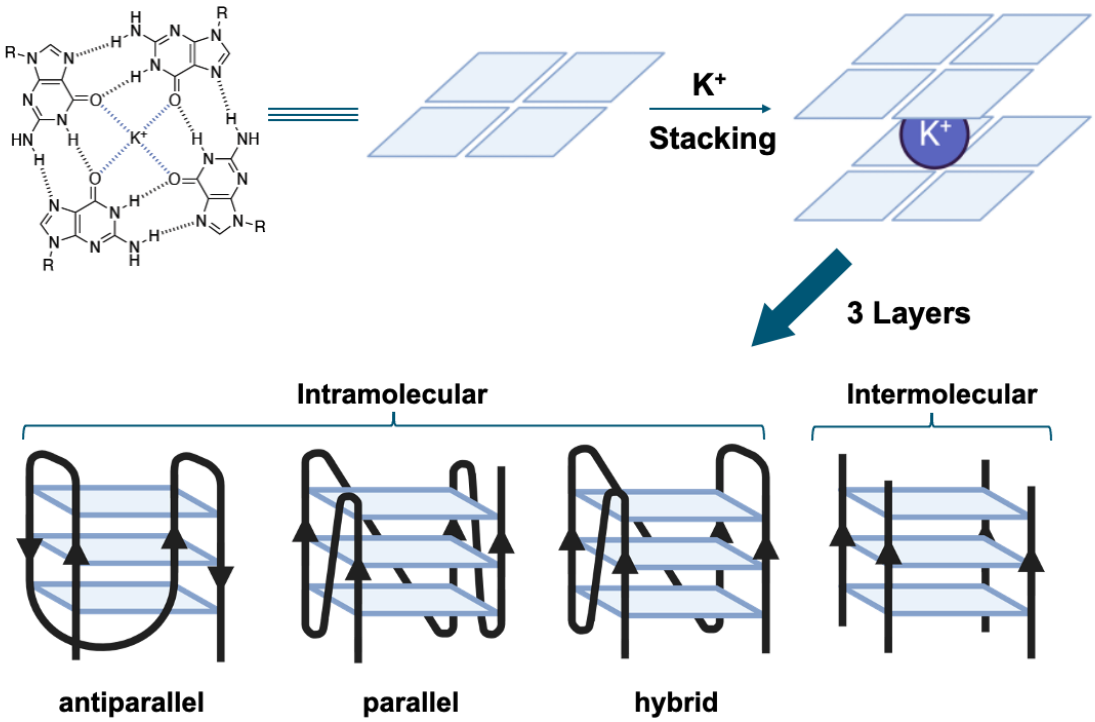

Figure 1.1. Hoogsteen hydrogen bonding together with coordinated cation brings four guanines together to form guanine quartet. Further stacking of at least three layers of G-quartet gives rise to G-quadruplex structure. G-quadruplex can form into various topologies and between multiple molecules.

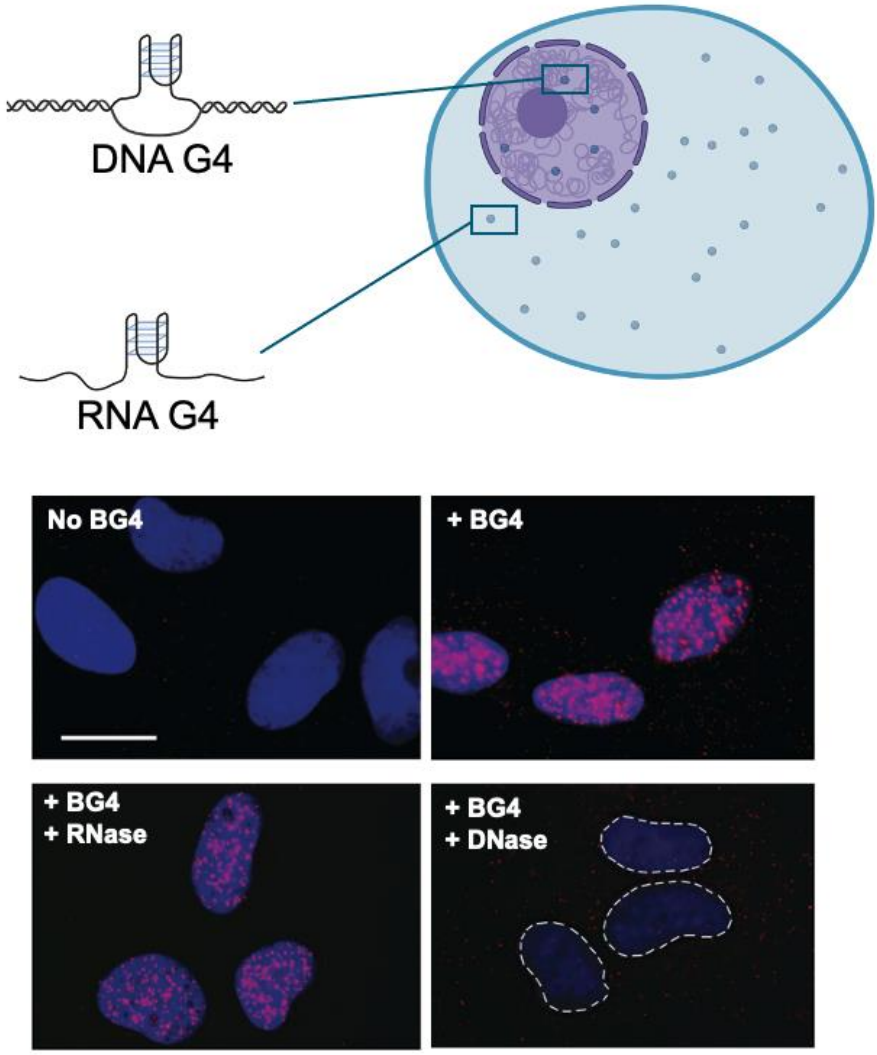

Figure 1.2. Visualization of G4 structures with a G4-specific antibody: DNA G4 structures are predominantly localized in the nucleus, while RNA G4 structures exhibit preferential cytoplasmic localization. Adopted from Ref. (141).

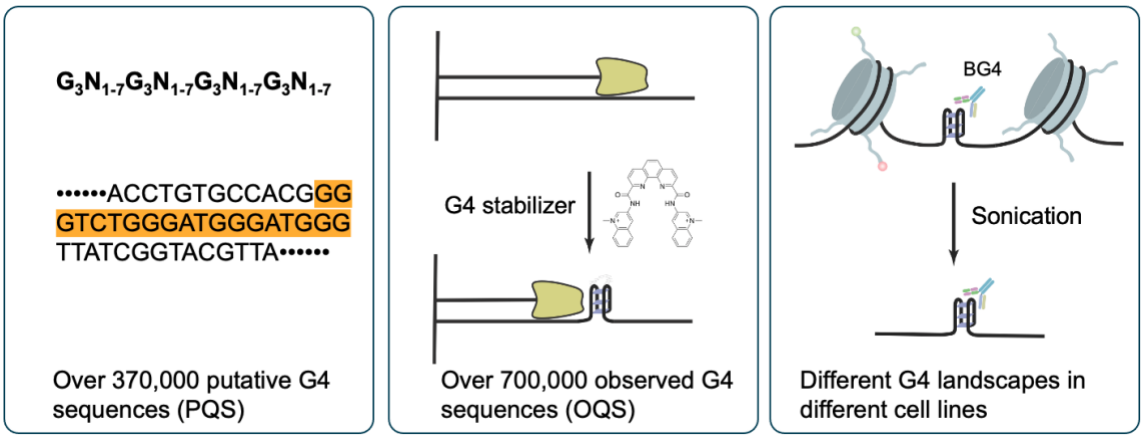

Figure 1.3. A schematic overview of genome-wide mapping techniques for DNA G4 structures, including consensus motif analysis, polymerase stalling (G4-seq), and G4-specific antibody immunoprecipitation (G4 ChIP-seq).

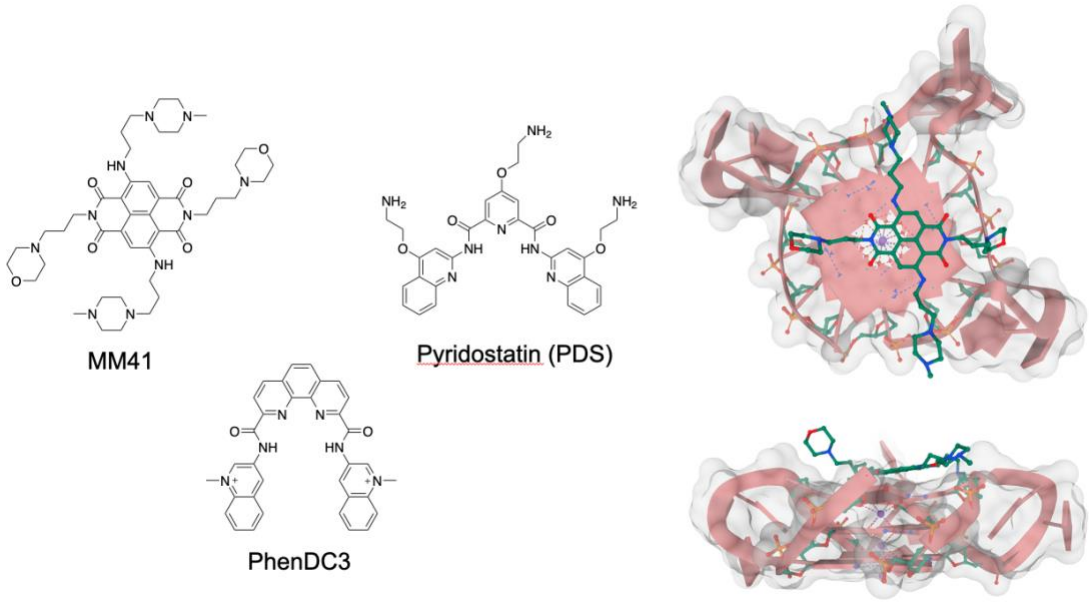

Figure 1.4. Chemical structures of small molecular G4-binding ligands including MM41, pyridostatin and PhenDC3. The crystal structure of MM41 in complex with DNA G4 shows $\boldsymbol{\pi}$-$\boldsymbol{\pi}$ stacking interactions between aromatic rings.

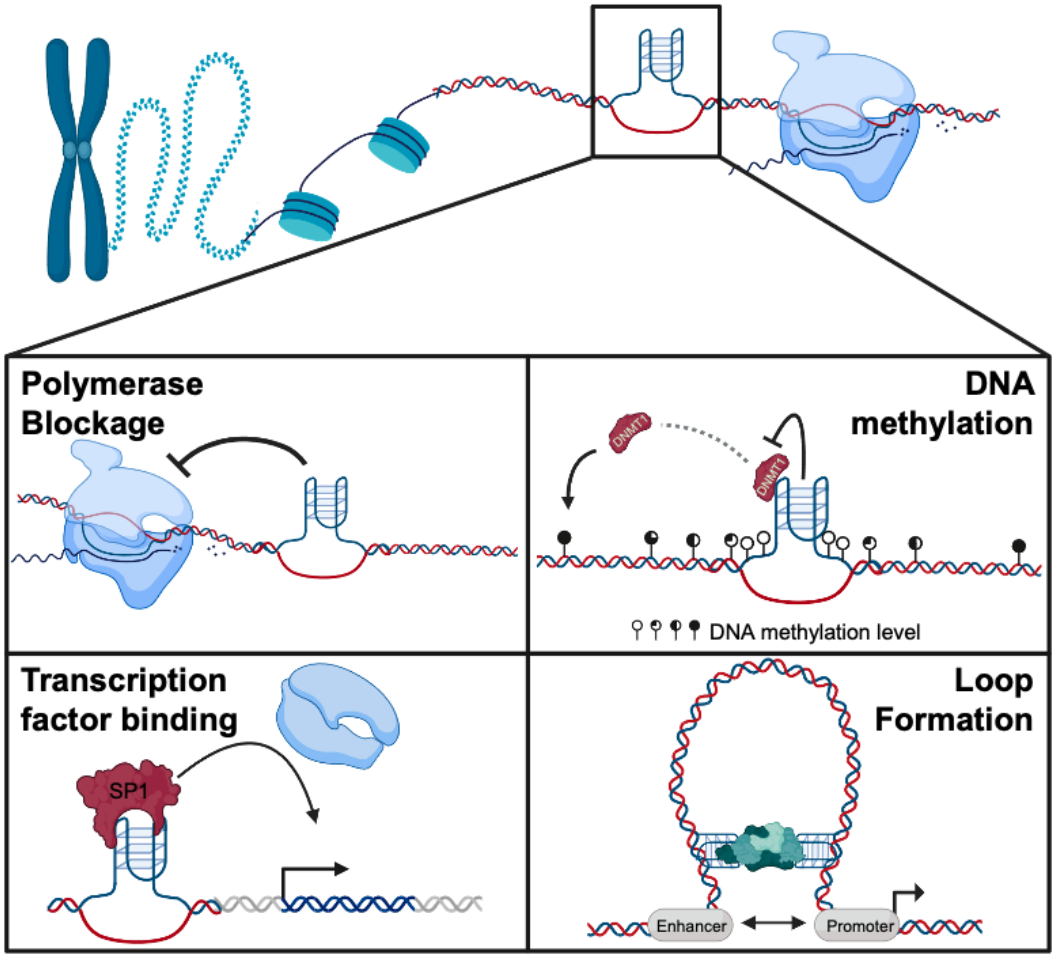

Figure 1.5. A schematic diagram illustrating the role of DNA G4 in transcription.

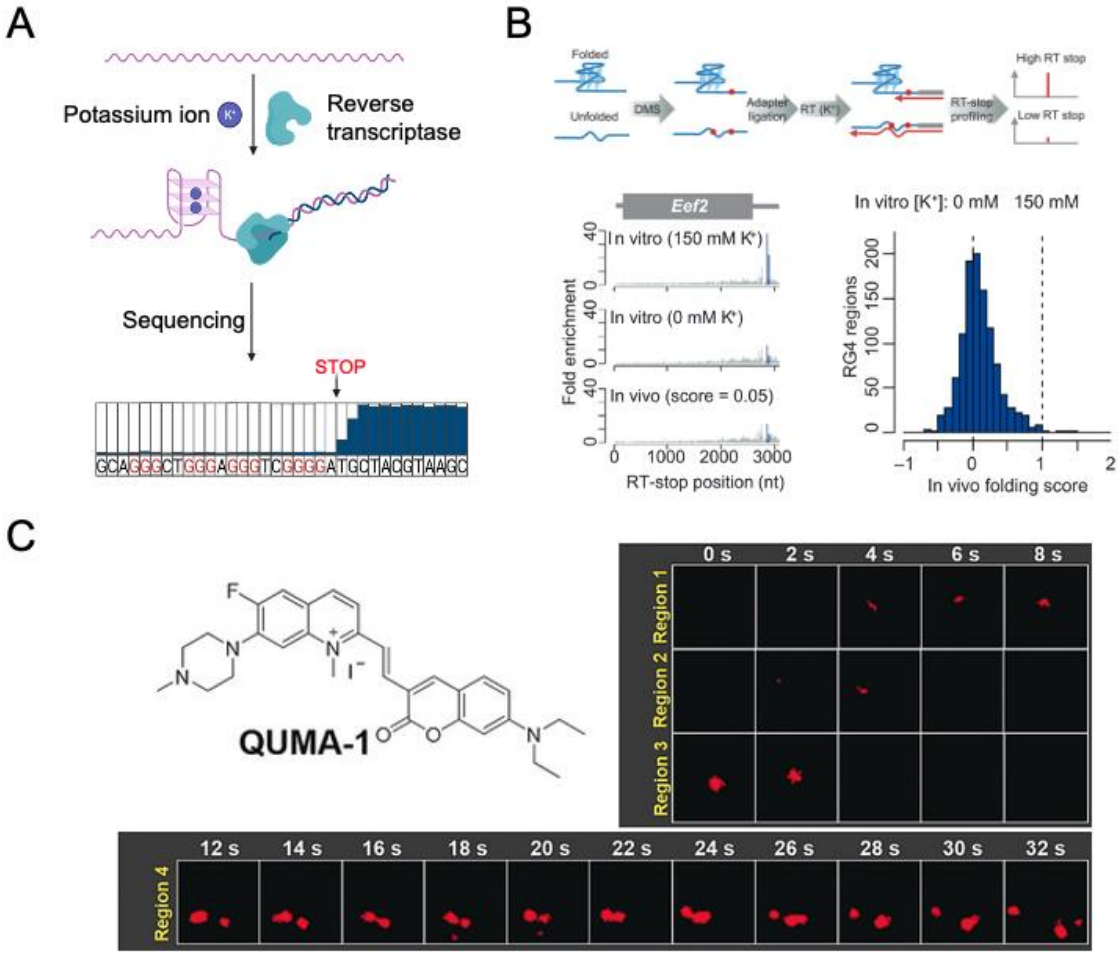

Figure 1.6. Skepticisms of intra-cellular RNA G4 formation. (A) rG4-seq detects RNA G4 through polymerase stalling. (B) Dimethyl sulfate-treatment sequencing indicates a predominantly unfolded state of RNA G4 in human cells. (C) The RNA G4-specific fluorescent probe, QUMA-1, reveals highly dynamic RNA G4 foci in live cells. Adopted from Ref. (142-144).

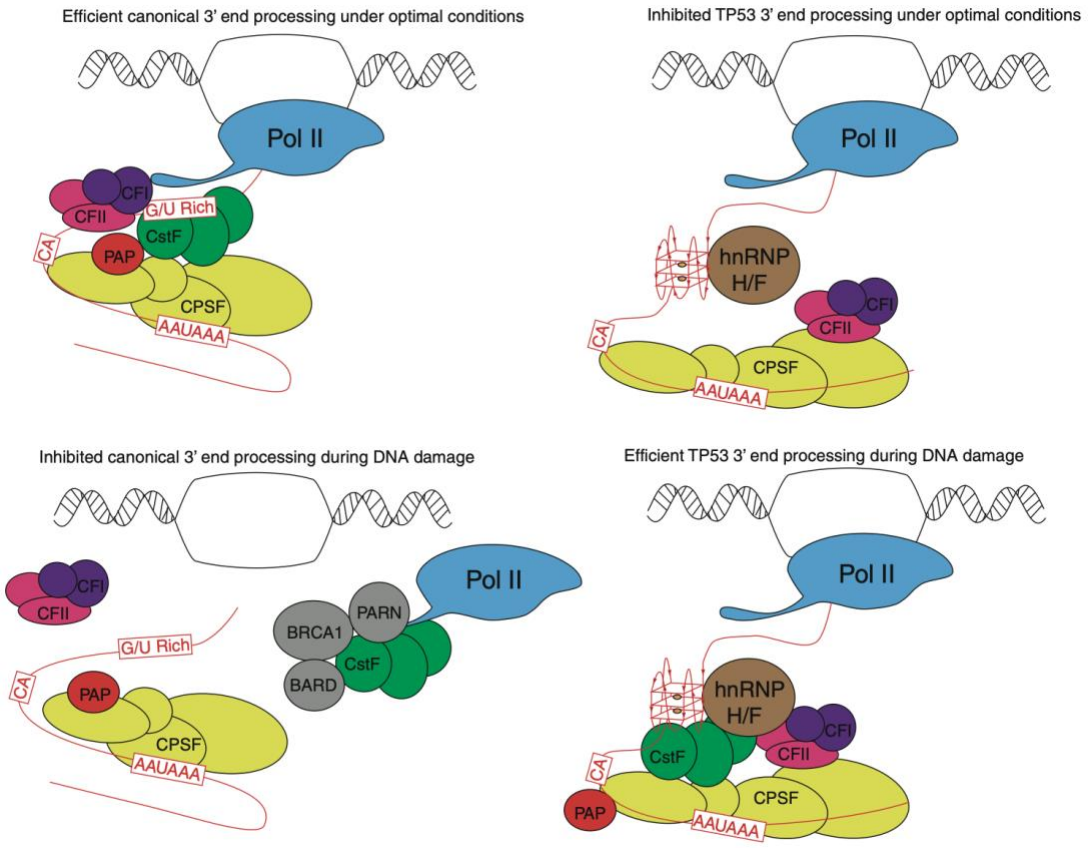

Figure 1.7. Proposed roles of RNA G4 structures in 3'-end mRNA processing. Adopted from Ref. (145).

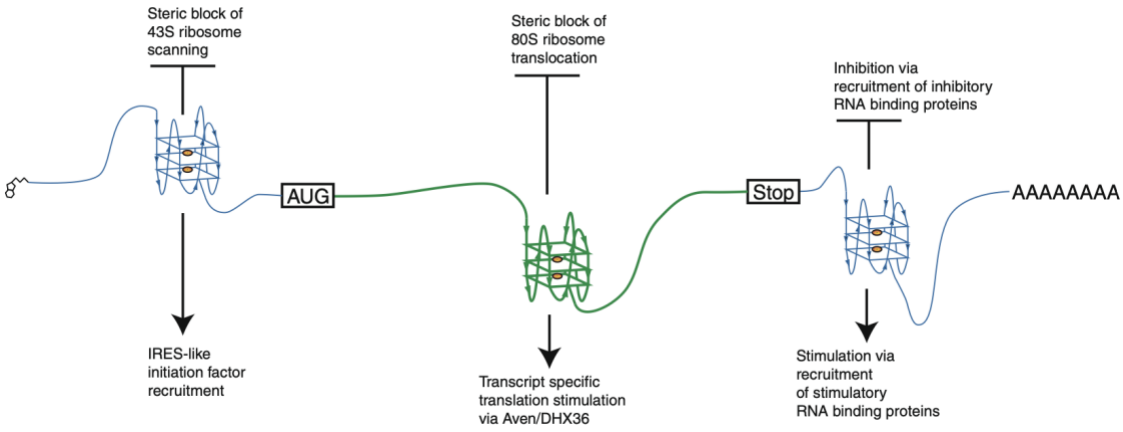

Figure 1.8. Proposed roles of RNA G4 structures in translation. Adopted from Ref. (145).

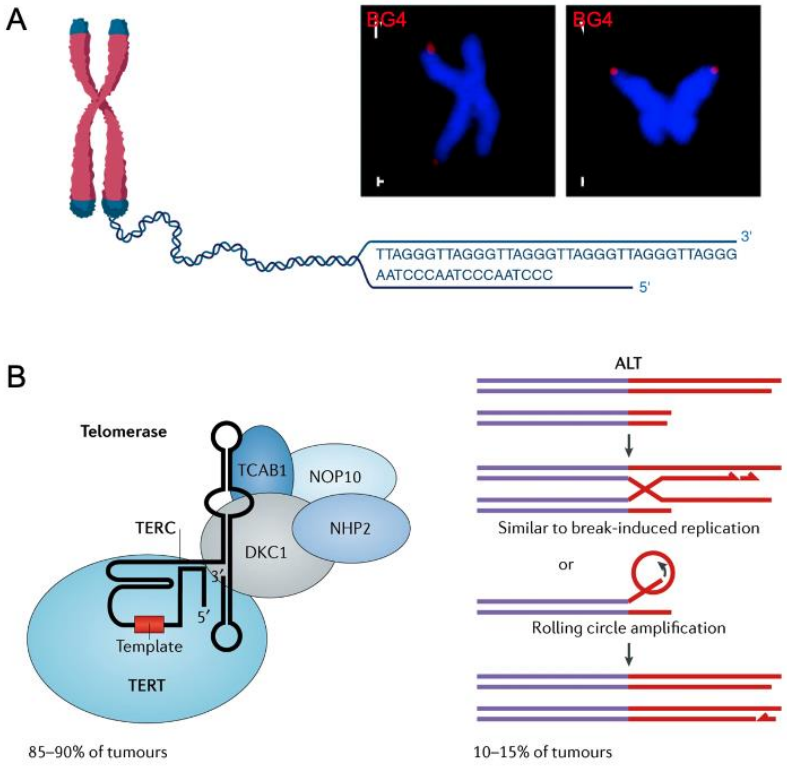

Figure 1.9. Structures and maintenance mechanisms of telomeres. (A) Composition of telomere DNA: TTAGGG double-stranded repeats accompanied by a single-stranded overhang. (B) Two predominant telomere maintenance mechanisms found in immortal cells, including tumor cells. Adopted from Refs. (20,146).

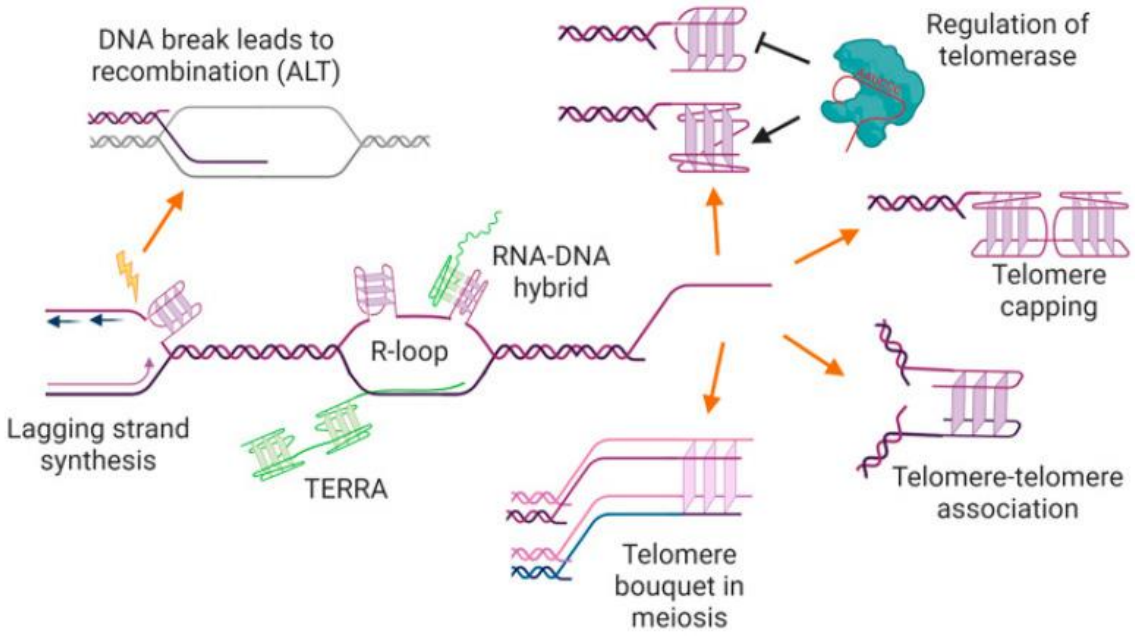

Figure 1.10. Potential locations, functions, and consequences of G4 structures at telomeres. Adopted from Ref. (147).

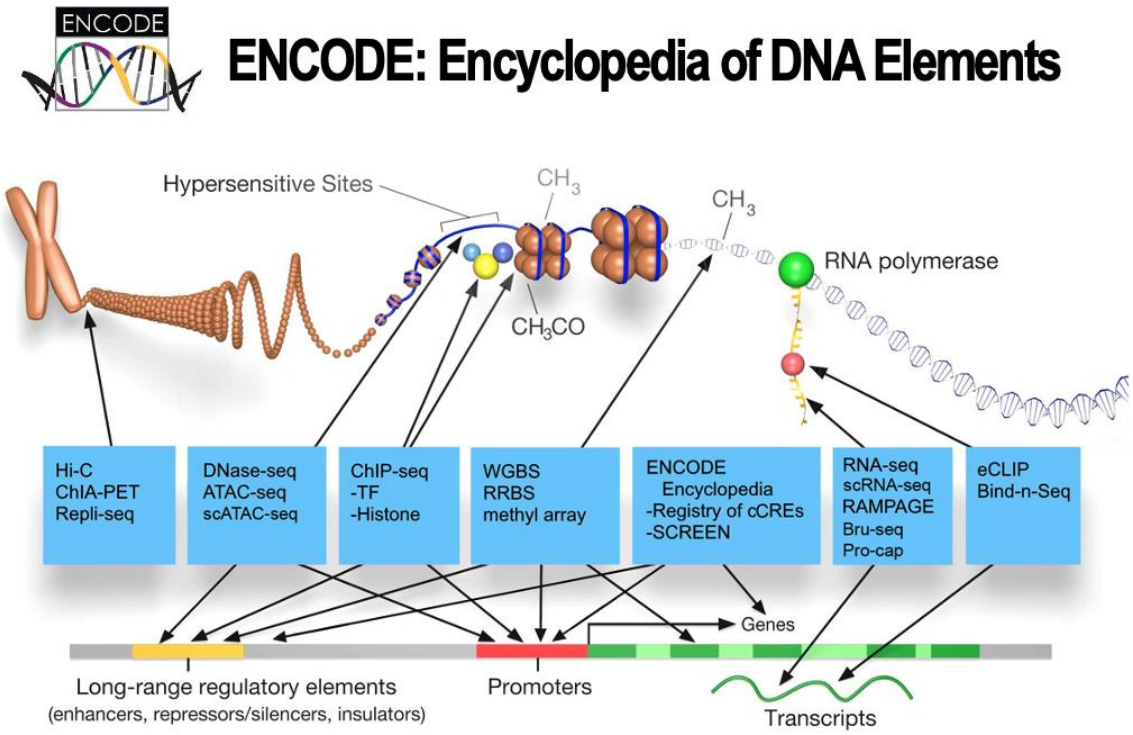

Figure 1.11. A schematic overview of ENCODE project, employing a variety of sequencing methods to identify functional elements. Adopted from Ref. (139).

1.7. References

1. A., T. and Muskhelishvili. (2015) DNA structure and function. *The FEBS journal*, **282**.

2. Bochman, M.L., Paeschke, K. and Zakian, V.A. (2012) DNA secondary structures: stability and function of G-quadruplex structures. *Nat. Rev. Genet.*, **13**, 770-780.

3. Burge, S., Parkinson, G.N., Hazel, P., Todd, A.K. and Neidle, S. (2006) Quadruplex DNA: sequence, topology and structure. *Nucleic Acids Res.*, **34**, 5402-5415.

4. Mukundan, V.T. and Phan, A.T. (2013) Bulges in G-quadruplexes: broadening the definition of G-quadruplex-forming sequences. *J. Am. Chem. Soc.*, **135**, 5017-5028.

5. Joachimi, A., Benz, A. and Hartig, J.S. (2009) A comparison of DNA and RNA quadruplex structures and stabilities. *Biorg. Med. Chem.*, **17**, 6811-6815.

6. Zhang, D.-H., Fujimoto, T., Saxena, S., Yu, H.-Q., Miyoshi, D. and Sugimoto, N. (2010) Monomorphic RNA G-quadruplex and polymorphic DNA G-quadruplex structures responding to cellular environmental factors. *Biochemistry*, **49**, 4554-4563.

7. Gellert, M., Lipsett, M.N. and Davies, D.R. (1962) Helix formation by guanylic acid. *Proceedings of the National Academy of Sciences*, **48**, 2013-2018.

8. Siddiqui-Jain, A., Grand, C.L., Bearss, D.J. and 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.

9. Parkinson, G.N., Lee, M.P. and Neidle, S. (2002) Crystal structure of parallel quadruplexes from human telomeric DNA. *Nature*, **417**, 876-880.

10. Patel, D.J., Phan, A.T. and Kuryavyi, V. (2007) Human telomere, oncogenic promoter and 5′-UTR G-quadruplexes: diverse higher order DNA and RNA targets for cancer therapeutics. *Nucleic Acids Res.*, **35**, 7429-7455.

11. Sun, D. and Hurley, L.H. (2010) Biochemical techniques for the characterization of G-quadruplex structures: EMSA, DMS footprinting, and DNA polymerase stop assay. *G-Quadruplex DNA: Methods and Protocols*, 65-79.

12. Han, H., Hurley, L.H. and Salazar, M. (1999) A DNA polymerase stop assay for G-quadruplex-interactive compounds. *Nucleic Acids Res.*, **27**, 537-542.

13. Fan, J.-H., Bochkareva, E., Bochkarev, A. and Gray, D.M. (2009) Circular dichroism spectra and electrophoretic mobility shift assays show that human replication protein A binds and melts intramolecular G-quadruplex structures. *Biochemistry*, **48**, 1099-1111.

14. Lin, C., Dickerhoff, J. and Yang, D. (2019) NMR studies of G-quadruplex structures and G-quadruplex-interactive compounds. *G-Quadruplex Nucleic Acids: Methods and Protocols*, 157-176.

15. Campbell, N.H. and Parkinson, G.N. (2007) Crystallographic studies of quadruplex nucleic acids. *Methods*, **43**, 252-263.

16. Huppert, J.L. and Balasubramanian, S. (2005) Prevalence of quadruplexes in the human genome. *Nucleic Acids Res.*, **33**, 2908-2916.

17. Chambers, V.S., Marsico, G., Boutell, J.M., Di Antonio, M., Smith, G.P. and Balasubramanian, S. (2015) High-throughput sequencing of DNA G-quadruplex structures in the human genome. *Nat. Biotechnol.*, **33**, 877-881.

18. Marsico, G., Chambers, V.S., Sahakyan, A.B., McCauley, P., Boutell, J.M., Antonio, M.D. and Balasubramanian, S. (2019) Whole genome experimental maps of DNA G-quadruplexes in multiple species. *Nucleic Acids Res.*, **47**, 3862-3874.

19. Kwok, C.K., Marsico, G., Sahakyan, A.B., Chambers, V.S. and Balasubramanian, S. (2016) rG4-seq reveals widespread formation of G-quadruplex structures in the human transcriptome. *Nat. Methods*, **13**, 841-844.

20. Biffi, G., Tannahill, D., McCafferty, J. and Balasubramanian, S. (2013) Quantitative visualization of DNA G-quadruplex structures in human cells. *Nat. Chem.*, **5**, 182-186.

21. Hansel-Hertsch, R., Beraldi, D., Lensing, S.V., Marsico, G., Zyner, K., Parry, A., Di Antonio, M., Pike, J., Kimura, H., Narita, M. *et al.* (2016) G-quadruplex structures mark human regulatory chromatin. *Nat. Genet.*, **48**, 1267-1272.

22. Hansel-Hertsch, R., Spiegel, J., Marsico, G., Tannahill, D. and Balasubramanian, S. (2018) Genome-wide mapping of endogenous G-quadruplex DNA structures by chromatin immunoprecipitation and high-throughput sequencing. *Nat. Protoc.*, **13**, 551-564.

23. Maizels, N. and Gray, L.T. (2013) The G4 genome. *PLoS Genet.*, **9**, e1003468.

24. Biffi, G., Di Antonio, M., Tannahill, D. and Balasubramanian, S. (2013) Visualization and selective chemical targeting of RNA G-quadruplex structures in the cytoplasm of human cells. *Nature Chemistry*, **6**, 75.

25. Mohanty, J., Barooah, N., Dhamodharan, V., Harikrishna, S., Pradeepkumar, P. and Bhasikuttan, A.C. (2013) Thioflavin T as an efficient inducer and selective fluorescent sensor for the human telomeric G-quadruplex DNA. *Journal of the American Chemical Society*, **135**, 367-376.

26. Monchaud, D., Allain, C. and Teulade-Fichou, M.-P. (2007) Thiazole orange: a useful probe for fluorescence sensing of G-quadruplex–ligand interactions. *Nucleosides, Nucleotides, and Nucleic Acids*, **26**, 1585-1588.

27. Spiegel, J., Adhikari, S. and Balasubramanian, S. (2020) The Structure and Function of DNA G-Quadruplexes. *Trends Chem*, **2**, 123-136.

28. Sun, D., Thompson, B., Cathers, B.E., Salazar, M., Kerwin, S.M., Trent, J.O., Jenkins, T.C., Neidle, S. and Hurley, L.H. (1997) Inhibition of human telomerase by a G-quadruplex-interactive compound. *J. Med. Chem.*, **40**, 2113-2116.

29. Marchetti, C., Zyner, K.G., Ohnmacht, S.A., Robson, M., Haider, S.M., Morton, J.P., Marsico, G., Vo, T., Laughlin-Toth, S. and Ahmed, A.A. (2018) Targeting multiple effector pathways in pancreatic ductal adenocarcinoma with a G-quadruplex-binding small molecule. *J. Med. Chem.*, **61**, 2500-2517.

30. Monchaud, D. and Teulade-Fichou, M.-P. (2008) A hitchhiker's guide to G-quadruplex ligands. *Org. Biomol. Chem.*, **6**, 627-636.

31. Clark, G.R., Pytel, P.D., Squire, C.J. and Neidle, S. (2003) Structure of the first parallel DNA quadruplex-drug complex. *J. Am. Chem. Soc.*, **125**, 4066-4067.

32. Liu, L.-Y., Ma, T.-Z., Zeng, Y.-L., Liu, W. and Mao, Z.-W. (2022) Structural basis of pyridostatin and its derivatives specifically binding to G-quadruplexes. *Journal of the American Chemical Society*, **144**, 11878-11887.

33. Li, Q., Xiang, J.F., Yang, Q.F., Sun, H.X., Guan, A.J. and Tang, Y.L. (2013) G4LDB: a database for discovering and studying G-quadruplex ligands. *Nucleic Acids Res.*, **41**, D1115-1123.

34. Capra, J.A., Paeschke, K., Singh, M. and Zakian, V.A. (2010) G-quadruplex DNA sequences are evolutionarily conserved and associated with distinct genomic features in Saccharomyces cerevisiae. *PLoS computational biology*, **6**, e1000861.

35. Lam, E.Y.N., Beraldi, D., Tannahill, D. and Balasubramanian, S. (2013) G-quadruplex structures are stable and detectable in human genomic DNA. *Nat. Commun.*, **4**, 1796.

36. Williams, J.D., Fleetwood, S., Berroyer, A., Kim, N. and Larson, E.D. (2015) Sites of instability in the human TCF3 (E2A) gene adopt G-quadruplex DNA structures in vitro. *Frontiers in genetics*, **6**, 177.

37. Kumari, R., Nambiar, M., Shanbagh, S. and Raghavan, S.C. (2015) Detection of G-quadruplex DNA using primer extension as a tool. *PLoS One*, **10**, e0119722.

38. Cheung, I., Schertzer, M., Rose, A. and Lansdorp, P.M. (2002) Disruption of dog-1 in Caenorhabditis elegans triggers deletions upstream of guanine-rich DNA. *Nat. Genet.*, **31**, 405-409.

39. Belotserkovskii, B.P., Liu, R., Tornaletti, S., Krasilnikova, M.M., Mirkin, S.M. and Hanawalt, P.C. (2010) Mechanisms and implications of transcription blockage by guanine-rich DNA sequences. *Proceedings of the National Academy of Sciences*, **107**, 12816-12821.

40. Gilmour, D.S. (2009) Promoter proximal pausing on genes in metazoans. *Chromosoma*, **118**, 1-10.

41. Eddy, J., Vallur, A.C., Varma, S., Liu, H., Reinhold, W.C., Pommier, Y. and Maizels, N. (2011) G4 motifs correlate with promoter-proximal transcriptional pausing in human genes. *Nucleic Acids Res.*, **39**, 4975-4983.

42. Belotserkovskii, B.P., Tornaletti, S., D'Souza, A.D. and Hanawalt, P.C. (2018) R-loop generation during transcription: Formation, processing and cellular outcomes. *DNA Repair*, **71**, 69-81.

43. Huppert, J.L. (2008) Thermodynamic prediction of RNA–DNA duplex-forming regions in the human genome. *Mol. Biosyst.*, **4**, 686-691.

44. Lee, C.-Y., McNerney, C., Ma, K., Zhao, W., Wang, A. and Myong, S. (2020) R-loop induced G-quadruplex in non-template promotes transcription by successive R-loop formation. *Nat. Commun.*, **11**, 3392.

45. Vizcaíno, C., Mansilla, S. and Portugal, J. (2015) Sp1 transcription factor: A long-standing target in cancer chemotherapy. *Pharmacol. Ther.*, **152**, 111-124.

46. Raiber, E.A., Kranaster, R., Lam, E., Nikan, M. and Balasubramanian, S. (2012) A non-canonical DNA structure is a binding motif for the transcription factor SP1 in vitro. *Nucleic Acids Res.*, **40**, 1499-1508.

47. Membrino, A., Cogoi, S., Pedersen, E.B. and Xodo, L.E. (2011) G4-DNA formation in the HRAS promoter and rational design of decoy oligonucleotides for cancer therapy. *PLoS One*, **6**, e24421.

48. Cogoi, S., Paramasivam, M., Membrino, A., Yokoyama, K.K. and Xodo, L.E. (2010) The KRAS promoter responds to Myc-associated zinc finger and poly(ADP-ribose) polymerase 1 proteins, which recognize a critical quadruplex-forming GA-element. *J. Biol. Chem.*, **285**, 22003-22016.

49. Vlasenok, M., Levchenko, O., Basmanov, D., Klinov, D., Varizhuk, A. and Pozmogova, G. (2018) Data set on G4 DNA interactions with human proteins. *Data in brief*, **18**, 348-359.

50. Lamonica, J.M., Deng, W., Kadauke, S., Campbell, A.E., Gamsjaeger, R., Wang, H., Cheng, Y., Billin, A.N., Hardison, R.C. and Mackay, J.P. (2011) Bromodomain protein Brd3 associates with acetylated GATA1 to promote its chromatin occupancy at erythroid target genes. *Proceedings of the National Academy of Sciences*, **108**, E159-E168.

51. Pavlova, II, Tsvetkov, V.B., Isaakova, E.A., Severov, V.V., Khomyakova, E.A., Lacis, I.A., Lazarev, V.N., Lagarkova, M.A., Pozmogova, G.E. and Varizhuk, A.M. (2020) Transcription-facilitating histone chaperons interact with genomic and synthetic G4 structures. *Int. J. Biol. Macromol.*, **160**, 1144-1157.

52. Moore, L.D., Le, T. and Fan, G. (2013) DNA methylation and its basic function. *Neuropsychopharmacology*, **38**, 23-38.

53. Mao, S.Q., Ghanbarian, A.T., Spiegel, J., Martinez Cuesta, S., Beraldi, D., Di Antonio, M., Marsico, G., Hansel-Hertsch, R., Tannahill, D. and Balasubramanian, S. (2018) DNA G-quadruplex structures mold the DNA methylome. *Nat. Struct. Mol. Biol.*, **25**, 951-957.

54. Gonzalez, V., Guo, K., Hurley, L. and Sun, D. (2009) Identification and characterization of nucleolin as a c-myc G-quadruplex-binding protein. *J. Biol. Chem.*, **284**, 23622-23635.

55. Pagano, B., Margarucci, L., Zizza, P., Amato, J., Iaccarino, N., Cassiano, C., Salvati, E., Novellino, E., Biroccio, A., Casapullo, A. *et al.* (2015) Identification of novel interactors of human telomeric G-quadruplex DNA. *Chem. Commun. (Camb.)*, **51**, 2964-2967.

56. Williams, P., Li, L., Dong, X. and Wang, Y. (2017) Identification of SLIRP as a G Quadruplex-Binding Protein. *J. Am. Chem. Soc.*, **139**, 12426-12429.

57. Gray, L.T., Vallur, A.C., Eddy, J. and Maizels, N. (2014) G quadruplexes are genomewide targets of transcriptional helicases XPB and XPD. *Nat. Chem. Biol.*, **10**, 313-318.

58. Spiegel, J., Cuesta, S.M., Adhikari, S., Hansel-Hertsch, R., Tannahill, D. and Balasubramanian, S. (2021) G-quadruplexes are transcription factor binding hubs in human chromatin. *Genome Biol.*, **22**, 117.

59. Zyner, K.G., Mulhearn, D.S., Adhikari, S., Martinez Cuesta, S., Di Antonio, M., Erard, N., Hannon, G.J., Tannahill, D. and Balasubramanian, S. (2019) Genetic interactions of G-quadruplexes in humans. *Elife*, **8**.

60. Rowley, M.J. and Corces, V.G. (2018) Organizational principles of 3D genome architecture. *Nature Reviews Genetics*, **19**, 789-800.

61. Dogan, E.S. and Liu, C. (2018) Three-dimensional chromatin packing and positioning of plant genomes. *Nat Plants*, **4**, 521-529.

62. Schoenfelder, S. and Fraser, P. (2019) Long-range enhancer-promoter contacts in gene expression control. *Nat. Rev. Genet.*, **20**, 437-455.

63. Lettice, L.A., Heaney, S.J., Purdie, L.A., Li, L., de Beer, P., Oostra, B.A., Goode, D., Elgar, G., Hill, R.E. and de Graaff, E. (2003) A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum. Mol. Genet.*, **12**, 1725-1735.

64. Amano, T., Sagai, T., Tanabe, H., Mizushina, Y., Nakazawa, H. and Shiroishi, T. (2009) Chromosomal dynamics at the Shh locus: limb bud-specific differential regulation of competence and active transcription. *Dev. Cell*, **16**, 47-57.

65. Rao, S.S., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D. and Lander, E.S. (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, **159**, 1665-1680.

66. Fudenberg, G., Imakaev, M., Lu, C., Goloborodko, A., Abdennur, N. and Mirny, L.A. (2016) Formation of chromosomal domains by loop extrusion. *Cell Rep.*, **15**, 2038-2049.

67. Merkenschlager, M. and Odom, D.T. (2013) CTCF and cohesin: linking gene regulatory elements with their targets. *Cell*, **152**, 1285-1297.

68. Hou, Y., Li, F., Zhang, R., Li, S., Liu, H., Qin, Z.S. and Sun, X. (2019) Integrative characterization of G-Quadruplexes in the three-dimensional chromatin structure. *Epigenetics*, **14**, 894-911.

69. Li, L., Williams, P., Ren, W., Wang, M.Y., Gao, Z., Miao, W., Huang, M., Song, J. and Wang, Y. (2021) YY1 interacts with guanine quadruplexes to regulate DNA looping and gene expression. *Nat. Chem. Biol.*, **17**, 161-168.

70. Todd, A.K., Johnston, M. and Neidle, S. (2005) Highly prevalent putative quadruplex sequence motifs in human DNA. *Nucleic Acids Res.*, **33**, 2901-2907.

71. Guo, J.U. and Bartel, D.P. (2016) RNA G-quadruplexes are globally unfolded in eukaryotic cells and depleted in bacteria. *Science*, **353**.

72. Chen, X.C., Chen, S.B., Dai, J., Yuan, J.H., Ou, T.M., Huang, Z.S. and Tan, J.H. (2018) Tracking the Dynamic Folding and Unfolding of RNA G-Quadruplexes in Live Cells. *Angew Chem Int Ed Engl*, **57**, 4702-4706.

73. Yang, S.Y., Lejault, P., Chevrier, S., Boidot, R., Robertson, A.G., Wong, J.M.Y. and Monchaud, D. (2018) Transcriptome-wide identification of transient RNA G-quadruplexes in human cells. *Nat Commun*, **9**, 4730.

74. Booy, E., Meier, M., Okun, N., Novakowski, S., Xiong, S., Stetefeld, J. and McKenna, S. (2012) The RNA helicase RHAU (DHX36) unwinds a G4-quadruplex in human telomerase RNA and promotes the formation of the P1 helix template boundary. *Nucleic Acids Res.*, **40**, 4110-4124.

75. Vaughn, J.P., Creacy, S.D., Routh, E.D., Joyner-Butt, C., Jenkins, G.S., Pauli, S., Nagamine, Y. and Akman, S.A. (2005) The DEXH protein product of the DHX36 gene is the major source of tetramolecular quadruplex G4-DNA resolving activity in HeLa cell lysates. *J. Biol. Chem.*, **280**, 38117-38120.

76. Chen, M.C., Tippana, R., Demeshkina, N.A., Murat, P., Balasubramanian, S., Myong, S. and Ferre-D'Amare, A.R. (2018) Structural basis of G-quadruplex unfolding by the DEAH/RHA helicase DHX36. *Nature*, **558**, 465-469.

77. Tippana, R., Chen, M.C., Demeshkina, N.A., Ferré-D'Amaré, A.R. and Myong, S. (2019) RNA G-Quadruplex is Resolved by Repetitive and ATP Dependent Mechanism of DHX36. *Biophys. J.*, **116**, 503a.

78. Colgan, D.F. and Manley, J.L. (1997) Mechanism and regulation of mRNA polyadenylation. *Genes Dev.*, **11**, 2755-2766.

79. Mandel, C.R., Bai, Y. and Tong, L. (2008) Protein factors in pre-mRNA 3′ -end processing. *Cellular and Molecular Life Sciences*, **65**, 1099-1122.

80. Zarudnaya, M.I., Kolomiets, I.M., Potyahaylo, A.L. and Hovorun, D.M. (2003) Downstream elements of mammalian pre‐mRNA polyadenylation signals: primary, secondary and higher‐order structures. *Nucleic Acids Res.*, **31**, 1375-1386.

81. Bagga, P.S., Ford, L.P., Chen, F. and Wilusz, J. (1995) The G-rich auxiliary downstream element has distinct sequence and position requirements and mediates efficient 3′ end pre-mRNA processing through a trans-acting factor. *Nucleic Acids Res.*, **23**, 1625-1631.

82. Dalziel, M., Nunes, N.M. and Furger, A. (2007) Two G-rich regulatory elements located adjacent to and 440 nucleotides downstream of the core poly (A) site of the intronless melanocortin receptor 1 gene are critical for efficient 3′ end processing. *Molecular and cellular biology*, **27**, 1568-1580.

83. Decorsière, A., Cayrel, A., Vagner, S. and Millevoi, S. (2011) Essential role for the interaction between hnRNP H/F and a G quadruplex in maintaining p53 pre-mRNA 3′-end processing and function during DNA damage. *Genes Dev.*, **25**, 220-225.

84. Kelemen, O., Convertini, P., Zhang, Z., Wen, Y., Shen, M., Falaleeva, M. and Stamm, S. (2013) Function of alternative splicing. *Gene*, **514**, 1-30.

85. Suhl, J.A., Chopra, P., Anderson, B.R., Bassell, G.J. and Warren, S.T. (2014) Analysis of FMRP mRNA target datasets reveals highly associated mRNAs mediated by G-quadruplex structures formed via clustered WGGA sequences. *Hum. Mol. Genet.*, **23**, 5479-5491.

86. Didiot, M.-C., Tian, Z., Schaeffer, C., Subramanian, M., Mandel, J.-L. and Moine, H. (2008) The G-quartet containing FMRP binding site in FMR1 mRNA is a potent exonic splicing enhancer. *Nucleic Acids Res.*, **36**, 4902-4912.

87. Gomez, D., Lemarteleur, T., Lacroix, L., Mailliet, P., Mergny, J.L. and Riou, J.F. (2004) Telomerase downregulation induced by the G‐quadruplex ligand 12459 in A549 cells is mediated by hTERT RNA alternative splicing. *Nucleic Acids Res.*, **32**, 371-379.

88. Sonenberg, N. and Hinnebusch, A.G. (2009) Regulation of translation initiation in eukaryotes: mechanisms and biological targets. *Cell*, **136**, 731-745.

89. Bugaut, A. and Balasubramanian, S. (2012) 5′-UTR RNA G-quadruplexes: translation regulation and targeting. *Nucleic Acids Res.*, **40**, 4727-4741.

90. Halder, K., Wieland, M. and Hartig, J.S. (2009) Predictable suppression of gene expression by 5′-UTR-based RNA quadruplexes. *Nucleic Acids Res.*, **37**, 6811-6817.

91. Morris, M.J., Negishi, Y., Pazsint, C., Schonhoft, J.D. and Basu, S. (2010) An RNA G-quadruplex is essential for cap-independent translation initiation in human VEGF IRES. *Journal of the American Chemical Society*, **132**, 17831-17839.

92. Li, W., Thakor, N., Xu, E.Y., Huang, Y., Chen, C., Yu, R., Holcik, M. and Kong, A.-N. (2010) An internal ribosomal entry site mediates redox-sensitive translation of Nrf2. *Nucleic Acids Res.*, **38**, 778-788.

93. Lee, S.C., Zhang, J., Strom, J., Yang, D., Dinh, T.N., Kappeler, K. and Chen, Q.M. (2016) G-Quadruplex in the NRF2 mRNA 5′ untranslated region regulates de novo NRF2 protein translation under oxidative stress. *Molecular and Cellular Biology*.

94. Arora, A. and Suess, B. (2011) An RNA G-quadruplex in the 3'UTR of the proto-oncogene PIM1 represses translation. *RNA Biol.*, **8**, 802-805.

95. Murat, P., Zhong, J., Lekieffre, L., Cowieson, N.P., Clancy, J.L., Preiss, T., Balasubramanian, S., Khanna, R. and Tellam, J. (2014) G-quadruplexes regulate Epstein-Barr virus–encoded nuclear antigen 1 mRNA translation. *Nat. Chem. Biol.*, **10**, 358-364.

96. Greider, C.W. (1991) Telomeres. *Curr. Opin. Cell Biol.*, **3**, 444-451.

97. de Lange, T. (2005) Shelterin: the protein complex that shapes and safeguards human telomeres. *Genes Dev*, **19**, 2100-2110.

98. Srinivas, N., Rachakonda, S. and Kumar, R. (2020) Telomeres and telomere length: a general overview. *Cancers (Basel)*, **12**, 558.

99. Shay, J.W. (2016) Role of telomeres and telomerase in aging and cancer. *Cancer Discov.*, **6**, 584-593.

100. Harley, C.B., Futcher, A.B. and Greider, C.W. (1990) Telomeres shorten during ageing of human fibroblasts. *Nature*, **345**, 458-460.

101. Maciejowski, J. and de Lange, T. (2017) Telomeres in cancer: tumour suppression and genome instability. *Nature reviews Molecular cell biology*, **18**, 175-186.

102. Kim, N.W., Piatyszek, M.A., Prowse, K.R., Harley, C.B., West, M.D., Ho, P.L., Coviello, G.M., Wright, W.E., Weinrich, S.L. and Shay, J.W. (1994) Specific association of human telomerase activity with immortal cells and cancer. *Science*, **266**, 2011-2015.

103. Shay, J.W. and Bacchetti, S. (1997) A survey of telomerase activity in human cancer. *Eur. J. Cancer*, **33**, 787-791.

104. Cong, Y.-S., Wright, W.E. and Shay, J.W. (2002) Human telomerase and its regulation. *Microbiology and molecular biology reviews*, **66**, 407-425.

105. Cesare, A.J. and Reddel, R.R. (2010) Alternative lengthening of telomeres: models, mechanisms and implications. *Nat. Rev. Genet.*, **11**, 319-330.

106. Bryan, T.M., Englezou, A., Gupta, J., Bacchetti, S. and Reddel, R.R. (1995) Telomere elongation in immortal human cells without detectable telomerase activity. *The EMBO journal*, **14**, 4240-4248.

107. Heaphy, C.M., De Wilde, R.F., Jiao, Y., Klein, A.P., Edil, B.H., Shi, C., Bettegowda, C., Rodriguez, F.J., Eberhart, C.G. and Hebbar, S. (2011) Altered telomeres in tumors with ATRX and DAXX mutations. *Science*, **333**, 425-425.

108. Goldberg, A.D., Banaszynski, L.A., Noh, K.-M., Lewis, P.W., Elsaesser, S.J., Stadler, S., Dewell, S., Law, M., Guo, X. and Li, X. (2010) Distinct factors control histone variant H3. 3 localization at specific genomic regions. *Cell*, **140**, 678-691.

109. Voon, H.P. and Wong, L.H. (2016) New players in heterochromatin silencing: histone variant H3. 3 and the ATRX/DAXX chaperone. *Nucleic Acids Res.*, **44**, 1496-1501.

110. Clatterbuck Soper, S.F. and Meltzer, P.S. (2023) ATRX/DAXX: Guarding the Genome against the Hazards of ALT. *Genes*, **14**, 790.

111. Henderson, E., Hardin, C.C., Walk, S.K., Tinoco, I., Jr. and Blackburn, E.H. (1987) Telomeric DNA oligonucleotides form novel intramolecular structures containing guanine-guanine base pairs. *Cell*, **51**, 899-908.

112. Tran, P.L.T., Mergny, J.-L. and Alberti, P. (2011) Stability of telomeric G-quadruplexes. *Nucleic Acids Res.*, **39**, 3282-3294.

113. Granotier, C., Pennarun, G., Riou, L., Hoffschir, F., Gauthier, L.R., De Cian, A., Gomez, D., Mandine, E., Riou, J.-F. and Mergny, J.-L. (2005) Preferential binding of a G-quadruplex ligand to human chromosome ends. *Nucleic Acids Res.*, **33**, 4182-4190.

114. Smith, J.S., Chen, Q., Yatsunyk, L.A., Nicoludis, J.M., Garcia, M.S., Kranaster, R., Balasubramanian, S., Monchaud, D., Teulade-Fichou, M.P., Abramowitz, L. *et al.* (2011) Rudimentary G-quadruplex-based telomere capping in Saccharomyces cerevisiae. *Nat. Struct. Mol. Biol.*, **18**, 478-485.

115. Jurikova, K., Gajarsky, M., Hajikazemi, M., Nosek, J., Prochazkova, K., Paeschke, K., Trantirek, L. and Tomaska, L. (2020) Role of folding kinetics of secondary structures in telomeric G-overhangs in the regulation of telomere maintenance in Saccharomyces cerevisiae. *J. Biol. Chem.*, **295**, 8958-8971.

116. Zimmer, J., Tacconi, E.M., Folio, C., Badie, S., Porru, M., Klare, K., Tumiati, M., Markkanen, E., Halder, S. and Ryan, A. (2016) Targeting BRCA1 and BRCA2 deficiencies with G-quadruplex-interacting compounds. *Mol. Cell*, **61**, 449-460.

117. Drosopoulos, W.C., Kosiyatrakul, S.T., Yan, Z., Calderano, S.G. and Schildkraut, C.L. (2012) Human telomeres replicate using chromosome-specific, rather than universal, replication programs. *J. Cell Biol.*, **197**, 253-266.

118. Paeschke, K., McDonald, K.R. and Zakian, V.A. (2010) Telomeres: structures in need of unwinding. *FEBS Lett.*, **584**, 3760-3772.

119. Drosopoulos, W.C., Kosiyatrakul, S.T. and Schildkraut, C.L. (2015) BLM helicase facilitates telomere replication during leading strand synthesis of telomeres. *J. Cell Biol.*, **210**, 191-208.

120. Wu, W., Bhowmick, R., Vogel, I., Özer, Ö., Ghisays, F., Thakur, R.S., Sanchez de Leon, E., Richter, P.H., Ren, L. and Petrini, J.H. (2020) RTEL1 suppresses G-quadruplex-associated R-loops at difficult-to-replicate loci in the human genome. *Nat. Struct. Mol. Biol.*, **27**, 424-437.

121. Luke, B. and Lingner, J. (2009) TERRA: telomeric repeat‐containing RNA. *The EMBO journal*, **28**, 2503-2510.

122. Takahama, K., Takada, A., Tada, S., Shimizu, M., Sayama, K., Kurokawa, R. and Oyoshi, T. (2013) Regulation of telomere length by G-quadruplex telomere DNA- and TERRA-binding protein TLS/FUS. *Chem Biol*, **20**, 341-350.

123. Takahama, K., Miyawaki, A., Shitara, T., Mitsuya, K., Morikawa, M., Hagihara, M., Kino, K., Yamamoto, A. and Oyoshi, T. (2015) G-quadruplex DNA-and RNA-specific-binding proteins engineered from the RGG domain of TLS/FUS. *ACS Chem. Biol.*, **10**, 2564-2569.

124. Sanger, F., Nicklen, S. and Coulson, A.R. (1977) DNA sequencing with chain-terminating inhibitors. *Proceedings of the national academy of sciences*, **74**, 5463-5467.

125. Sanger, F., Air, G.M., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddes, J.C., Hutchison III, C.A., Slocombe, P.M. and Smith, M. (1977) Nucleotide sequence of bacteriophage φX174 DNA. *Nature*, **265**, 687-695.

126. Consortium, I.H.G.S. (2004) Finishing the euchromatic sequence of the human genome. *Nature*, **431**, 931-945.

127. Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.-J. and Chen, Z. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376-380.

128. Furey, T.S. (2012) ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. *Nat. Rev. Genet.*, **13**, 840-852.

129. Hentze, M.W., Castello, A., Schwarzl, T. and Preiss, T. (2018) A brave new world of RNA-binding proteins. *Nature reviews Molecular cell biology*, **19**, 327-341.

130. Hafner, M., Katsantoni, M., Köster, T., Marks, J., Mukherjee, J., Staiger, D., Ule, J. and Zavolan, M. (2021) CLIP and complementary methods. *Nature Reviews Methods Primers*, **1**, 20.

131. Van Nostrand, E.L., Pratt, G.A., Shishkin, A.A., Gelboin-Burkhart, C., Fang, M.Y., Sundararaman, B., Blue, S.M., Nguyen, T.B., Surka, C. and Elkins, K. (2016) Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat. Methods*, **13**, 508-514.

132. Ascano, M., Hafner, M., Cekan, P., Gerstberger, S. and Tuschl, T. (2012) Identification of RNA–protein interaction networks using PAR‐CLIP. *Wiley Interdisciplinary Reviews: RNA*, **3**, 159-177.

133. Dominguez, C., Fisette, J.-F., Chabot, B. and Allain, F.H.T. (2010) Structural basis of G-tract recognition and encaging by hnRNP F quasi-RRMs. *Nature Structural & Molecular Biology*, **17**, 853-861.

134. Bonev, B. and Cavalli, G. (2016) Organization and function of the 3D genome. *Nat. Rev. Genet.*, **17**, 661-678.

135. Kempfer, R. and Pombo, A. (2020) Methods for mapping 3D chromosome architecture. *Nat. Rev. Genet.*, **21**, 207-226.

136. Mumbach, M.R., Rubin, A.J., Flynn, R.A., Dai, C., Khavari, P.A., Greenleaf, W.J. and Chang, H.Y. (2016) HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nat. Methods*, **13**, 919-922.

137. Fang, R., Yu, M., Li, G., Chee, S., Liu, T., Schmitt, A.D. and Ren, B. (2016) Mapping of long-range chromatin interactions by proximity ligation-assisted ChIP-seq. *Cell Res.*, **26**, 1345-1348.

138. Consortium, E.P. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57-74.

139. Davis, C.A., Hitz, B.C., Sloan, C.A., Chan, E.T., Davidson, J.M., Gabdank, I., Hilton, J.A., Jain, K., Baymuradov, U.K., Narayanan, A.K. *et al.* (2018) The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res.*, **46**, D794-D801.

140. Landt, S.G., Marinov, G.K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., Bernstein, B.E., Bickel, P., Brown, J.B., Cayting, P. *et al.* (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.*, **22**, 1813-1831.

141. Biffi, G., Di Antonio, M., Tannahill, D. and Balasubramanian, S. (2014) Visualization and selective chemical targeting of RNA G-quadruplex structures in the cytoplasm of human cells. *Nat. Chem.*, **6**, 75-80.

142. Kwok, C.K., Marsico, G., Sahakyan, A.B., Chambers, V.S. and Balasubramanian, S. (2016) rG4-seq reveals widespread formation of G-quadruplex structures in the human transcriptome. *Nat. Methods*, **13**, 841-844.

143. Guo, J.U. and Bartel, D.P. (2016) RNA G-quadruplexes are globally unfolded in eukaryotic cells and depleted in bacteria. *Science*, **353**, aaf5371.

144. Chen, X.C., Chen, S.B., Dai, J., Yuan, J.H., Ou, T.M., Huang, Z.S. and Tan, J.H. (2018) Tracking the dynamic folding and unfolding of RNA G‐quadruplexes in live cells. *Angew. Chem.*, **130**, 4792-4796.

145. Fay, M.M., Lyons, S.M. and Ivanov, P. (2017) RNA G-quadruplexes in biology: principles and molecular mechanisms. *J. Mol. Biol.*, **429**, 2127-2147.

146. Shay, J.W. and Wright, W.E. (2019) Telomeres and telomerase: three decades of progress. *Nature Reviews Genetics*, **20**, 299-309.

147. Bryan, T.M. (2020) G-Quadruplexes at Telomeres: Friend or Foe? *Molecules*, **25**.

Chapter 2: Epigenetic functions of DNA G-quadruplexes on transcription: An Overlapping Calculation Approach

2.1. Introduction

The DNA guanine quadruplexes (G4) are four-stranded nucleic acid secondary structures folded in guanine-rich regions of DNA (1,2). Hoogsteen hydrogen bonding together with a monovalent cation, in the order $K^+ > Na^+ > Li^+$, brings four guanines together to form one layer of guanine quartet structure and further stacking of at least three layers of G quartet gives rise to G4. Cellular studied showed large variations between G4 structures including inter- or intra-strand, different loop size, different topology and even bulges because of discontinuities of Gs (3).

Different methods have been developed to detect and map G4 structures. Computational prediction using the consensus motif of $G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}N_{1-7}$ identified over 370,000 putative G4 sequences (PQS) in the human genome, where most of them are enriched in genome regulatory regions including telomeres, promoters and 5'-untranslated regions (4). The development of G4-structure specific antibody (BG4) enabled probing the presence of native G4s structures in cells (5). BG4 based ChIP-seq (G4 ChIP-seq) revealed landscape of G4s in various cell lines (6). Different cell-type exhibits a substantial number of distinct G4 sites, suggesting that G4 structures are cell type-specific and thus cell state-specific (7).

The discovery of G4s' prevalence in gene regulatory regions inspired research into their roles in transcriptional regulation. The association between endogenous G4 structures and DNA methylation, combined with the identification of DNMT1 as a G4-binding protein, suggests that G4 might play a crucial role in epigenetic regulation (8). Using publicly available ChIP-seq data for transcription factors in tandem with G4 ChIP-seq, Jochen et al. identified DNA G-

quadruplexes as critical binding hubs for various transcription factors, underscoring their importance in modulating gene expression (9).

Chromatin structure, nucleosome positioning and gene regulation are extensively related to histones and histone modifications (10). Aside from histone variants like H2AX and H2AZ, post-translational modifications of core histones are important in nucleosome structure remodeling. Lysine residues in histone tails can be methylated by SET-domain containing proteins and those in histone cores can be methylated by non-SET domain containing (11,12). Two families of histone demethylases, including amino oxidase homolog lysine demethylase 1 (KDM1) family and JmjC domain-containing family, are capable of removal of methylation (13).

Histone methylation can be accompanied with both transcriptional activation and repression depending on methylation site and state. H3K4me1 and H3K4me3 are activation marks, whereas H3K9me3 and H3K27me3 are related to transcriptional inhibition (14). Previous studies showed an association between G4 structure and open-chromatin structure (8). However, the role of G4 in histone modification and its interplay with histone-modifying proteins is understudied.

In this study, we first reproduced overlapping analysis using three different approaches to cross-validate our results. We extensively investigated the co-localization patterns of transcription factors within endogenous G4 structures, unveiling a complex interplay between TFs and G4. This interaction network pinpointed several protein clusters, indicating G4's involvement in key biological processes. By analyzing histone modification ChIP-seq datasets alongside BG4 ChIP-seq, we highlighted a potential synergy between DNA secondary structures and histone modifications.

2.2. Materials and Methods

Overlapping analysis

Genome-wide overlapping analyses of transcription factor binding sites in the ENCODE project with BG4 antibody-based ChIP-seq mapping of endogenous G-quadruplex structure *in vivo* (BG4 peaks) were performed (15). The ChIP-seq data for proteins were retrieved from the ENCODE portal under TF ChIP-seq and K562 cell line category. Extremely low read depth, un-replicated and drug treatment experiments were excluded. A total of 431 experimental results were downloaded and processed for overlapping analysis. G4 ChIP-seq data of K562 cell line were obtained using GEO accession number GSE107690. To ensure target proteins' peaks from true biological replicate, conservative irreproducible discovery rate (IDR) threshold peaks were used for overlapping analysis. IntervalStats were employed for overlapping percentage and p-value calculations (16). High-confidence peaks of G4 ChIP-seq data were used as reference while hg38 genome was used as domain. Conservative IDR thresholded peak file of each experiment was used as query file. Overlapping percentage was calculated as *overlapped peak number/total peak number of target protein*.

Network analyses

Jaccard similarity coefficient was employed to calculate the co-overlapping level between two datasets of interest. Correlated were filtered by identifying outliers based on co-overlapping level, that exceed 1.5 or 3.0, respectively, for moderate and strong specific partners, and an interquartile range above the 75^{th} percentile of co-overlapping levels (17). Network was generated using Gephi tool with all high confidence proteins that at least co-overlapped with one other protein (18). The color and weight of edges corresponds to co-overlapping levels between two target proteins. The radius of nodes represents the number of correlated co-overlapping proteins. The graph was partitioned into subnetworks using an algorithm developed by Bondel et al. and

implemented in Gephi, with options *randomize*, *use edge weights* and a resolution of 0.50 (19). ChIP-seq peak annotations were conducted using HOMER on the GRCh38 assembly utilizing the annotatePeaks.pl tool (20). Genes exhibiting peaks within their promoters were subjected to Gene Ontology analysis through DAVID (21,22).

Differential binding analysis

Differential ChIP-seq analysis was conducted using *Diffbind* software with default parameters (23). BG4 ChIP-seq in HepG2 and K562 were retrieved using GEO accession number GSE145090 and reprocessed in GRCh38 assembly (9). SP1 ChIP-seq in HepG2 and K562 were retrieved from ENCODE using accession number ENCSR460YAM and ENCSR372IML.

Enrichment profiling of histone modification

The enrichment profile analysis of histone modifications with BG4 peaks was obtained by using *bwtool* aggregate function (24). ChIP-seq signals were plotted within the +/- 2000 bps with respect to histone modification peak center or BG4 peak centers.

2.3. Results

Overlapping analysis

We first conducted genome-wide overlapping analyses of 431 target proteins from the ENCODE project with BG4 antibody-based ChIP-seq mapping of endogenous G-quadruplex structure *in vivo* (BG4 peaks) (Figure 2.1). Two different methods were employed: the n-bp cutoff and IntervalStats. The n-bp cutoff method is intuitive in which the intersection of two datasets is calculated and filtered based on chosen window size and two criteria were selected. 1-bp was chosen as a loose criterion, given its use in other studies (8,9). In this case, 1 bp was first selected as a loose criterion, where for a protein to be tagged as overlap with BG4 requires at least one base-pair of overlap for two binding sites. Considering that the consensus motif of G4 ($G_{3+}N_{1-}$

$_{7}G_{3+}N_{1-7}G_{3+}N_{1-7}$) spans approximately 30 bps, we also established a stricter threshold of 30 bps for overlapped tagging. On the other hand, IntervalStats examines each peak region from target protein dataset against BG4 peaks, calculating an exact p-value for every proximity event. The overlapping results from these three approaches yield consistent results. By combing the results, we obtained 105 high-confidence potential G4-interacting proteins (Figure 2.2).

Network analysis

Each potential G4-binding protein overlaps with certain BG4 binding area, i.e. a cellular G4 structure site. By comparing the overlapping patterns of these proteins of interest, we generated a similarity-based network that can reveal potential G4-interacting protein complex was generated (Figure 2.3). Gephi tool was used to create and visualize the network containing 105 co-overlapped target proteins. The weight of the edges connected two proteins represents co-overlapping specificity while the color indicates their co-overlapping level. Target proteins were shown to be highly correlated in terms of G-quadruplex overlapping while some clusters can be observed and labeled with different colors. Those clusters suggested multiple G4-binding protein complexes and their binding profiles. Subnetwork 5, for example, contains several proteins including FUS and NONO, which have been identified as G-quadruplex binding proteins (25,26). Gene ontology analysis showed multiple proteins (HNRNPL, HNRNPK, RBM15, FIP1L1, FUS, PCBP1, NONO, U2AF1 and RBM22) in this co-overlapping cluster are involved in mRNA splicing which indicates potential G-quadruplex functions in pre-mRNA processing (Figure 2.4). In contrast to previous observations that G-quadruplexes in the promoter regions are correlated with active transcription, subnetwork 3 is highly linked with negative regulation of transcription and histone deacetylation. HDAC1/2, GATAD2A are part of nucleosome remodeling and deacetylase (NuRD) complex which was known to repressively regulate downstream gene expression (27). Annotation of the co-overlapping BG4 peaks showed enrichment in cell division,

cell cycle control, protein degradation and DNA damage response pathway, suggestsing possible G4-dependent negative regulatory functions in those cellular activities (Figure 2.5).

Epigenetic crosslink between G4 and histone modifications

The detection of histone-modifying complexes among potential G4-interacting protein clusters hints at a potential link between G4 structures and histone modifications. To explore deeper into this prospective epigenetic interaction, we carried out an overlapping analysis focused specifically on histone modifications.

We obtained histone ChIP-seq datasets for K562 and HepG2 cells from the ENCODE project. A total of 34 experiments encompassing 12 distinct histone modifications was downloaded and processed for overlapping analysis with BG4-based ChIP-seq specific to each cell line (Table 2.1).

We observed high overlapping percentages in active histone marks, such as H3K4me3 (37.16% in K562 and 28.28% in HepG2), H3K9ac (26.05% in K562 and 24.23% in HepG2), and H3K27ac (22.47% in K562 and 21.12% in HepG2). These results substantiated the notion of a positive correlation between G4 structures and transcriptional activation. In contrast, repressive histone marks, like H3K9me3 (1.35% in K562 and 0.07% in HepG2) and H3K27me3 (0.36% in K562 and 0.45% in HepG2), exhibited extremely minimal colocalization. Enrichment profiles of H3K4me3, in relation to either G4 or H3K4me3 peak center, showcased a proximal co-occurrence with a crest-trough distance averaging 300 bp (Figure 2.6).

Interestingly, this crest-trough pattern differs from the overlap seen between transcription factors and G4 structures, where the peak of the transcription factor's binding site aligns precisely with G4 structures. Upon examining MNase-seq mapping of nucleosome positions, we verified that this shift arises due to nucleosome depletion within the G4 formation region, while

H3K4me3 enrichment appears in the neighboring genomic regions (28). A similar trend was also observed for H3K9ac and H3K27ac (Figure 2.7).

To explore further into the role of the G4 structure in histone methylation, we carried out overlapping analysis on histone-modifying enzymes, encompassing both histone methyltransferases and histone demethylases (Table 2.2). In line with the observed co-localization between G4 and H3K4me3, we found a high overlapping percentage involving HCFC1 and RBBP5 – both proteins play roles in H3K4 methyltransferase complexes. Notably, we also detected high overlapping percentages with enzymes responsible for histone demethylation, e.g., KDM5A (37.39% in HepG2) and KDM5B (36.40% in K562 and 29.22% in HepG2). A more detailed analysis of these overlapping regions revealed that HCFC1 and KDM5A in HepG2 cells, co-localized with endogenous G4 structures rather than being mutually exclusive (Figure 2.8). These findings hint at the dynamic role G4 structures play in the regulation of H3K4me3.

Regarding H3K9 methylation, SETDB1, which specifically trimethylates H3K9, displayed a reduced overlapping percentage when compared to enzymes related to H3K4me3 (29,30). This aligns with our earlier observations where H3K9me3 was depleted within G4 regions. Additionally, we observed a significantly higher overlapping percentage for the demethylase KDM4B, registering at 42.38% in K562 cells. This suggests that G4 structures might act as a scaffold for the recruitment of KDM4B, promoting the removal of H3K9 methylation and hence fostering a more active chromatin state.

2.4. Conclusions

Recent research into the role of G4 structures in transcription has revealed that G4s are prevalently located in gene regulatory regions, influencing transcriptional activity by either promoting or hindering the binding of transcriptional machinery (31). Their presence has been

linked to various biological processes, including epigenetic regulation, given their association with DNA methylation and specific binding proteins like DNMT1 (8). Additionally, studies leveraging ChIP-seq data have underscored G4s as crucial binding hubs for numerous transcription factors, emphasizing their significance in modulating gene expression and their potential interplay with other epigenetics marks, such as histone modifications (9).

Overlapping analysis serves as an intuitive method for preliminary assessment of interactions based on genomic proximity. Multiple important G4-interacting proteins were discovered based on binding site overlapping with putative G4 forming sequences. Telomere repeat-binding factor 2 (TRF2) ChIP-seq assay demonstrated genome-wide enrichment of G4 motifs and downstream biological experiment confirmed its G4-dependent epigenetic regulatory functions (32).

Transcription activation and epigenetic regulations are performed by protein complexes composed of multiple subunits (33). Analyzing overlaps based solely on individual proteins does not comprehensively depict all biologically significant proximal events. Furthermore, the crosslinking step in ChIP-seq experiments renders it challenging to differentiate direct interactions from indirect binding (34). In this study, by employing IntervalStats and n-bp based overlapping analyses, we first identified a set of potential G4-binding proteins. Comparing the G4 overlapping profiles of these candidates enabled us to construct a similarity-based network, which is partitioned into several protein clusters. Remarkably, each cluster displayed varied molecular functions, and some even exhibited opposing transcriptional repression activities. This suggests a previously unrecognized regulatory role of the G4 structure.

G-quadruplexes have also been recognized for their association with open chromatin structures and epigenetic regulations. Subramanian et al. discovered a pronounced correlation between endogenous G4 formation and DNA hypomethylation (8). Given the preferential binding

of DNA methyltransferase DNMT1 to G4 structures, they introduced a sequestration model wherein the G4 structure inhibits DNMT1 to modulate DNA methylation levels. In our research, we delved deeper into the relationship between the G4 structure and histone modifications. We provided evidence that active histone modifications, such as H3K4me3 and H3K9ac, show a strong association with endogenous G4 structures. In contrast, repressive histone marks, like H3K9me3, are less prevalent at G4 sites. By integrating overlapping analyses of histone methyltransferases and demethylases, our findings suggest a potential role for G4 in the regulation of histone modifications.

The current study possesses several limitations. Firstly, the reliability of high-affinity antibody-based mapping for G4 structures *in vivo* remains uncertain. The potential effects on G4 structure stability introduced by exogenous G4 antibodies aren't fully understood. There's a risk that artificial G4 structures, typically unfolded *in vivo*, might emerge. Secondly, the overlapping analysis is particularly sensitive to the quality of ChIP-seq datasets and the randomness inherent in protein-DNA binding activity. While the ENCODE project has implemented both antibody characterization and data analysis quality metrics, the mapping of protein binding sites might still be incomplete. Additionally, due to the absence of BG4 ChIP-seq experiments in various cell lines, only a fraction (specifically the K562 and HepG2 cell lines) of ENCODE's data was applicable. Datasets from a broader range of cell lines would not only offer a cross-validation of potential protein-G4 interactions, but also shed light on G4-dependent regulatory functions across diverse cell types.

In conclusion, leveraging ChIP-seq data from over 400 proteins sourced from the ENCODE project, we crafted an interaction network of potential G4-binding proteins. Coupled with histone modification mapping, this study offers fresh insights into the interplay between DNA secondary structures, epigenetic markers, and potentially involved cellular proteins. We

anticipate that our findings will provide a valuable foundation for experimental researchers to further explore G4-dependent gene regulatory mechanisms.

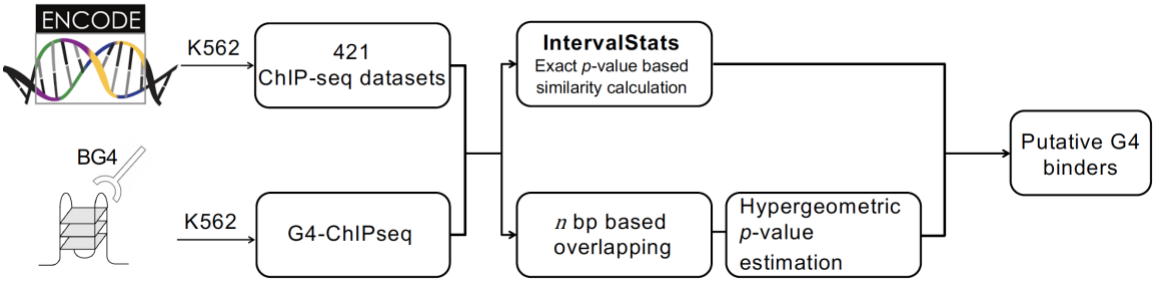

Figure 2.1. A schematic diagram showing the workflow of overlapping analysis.

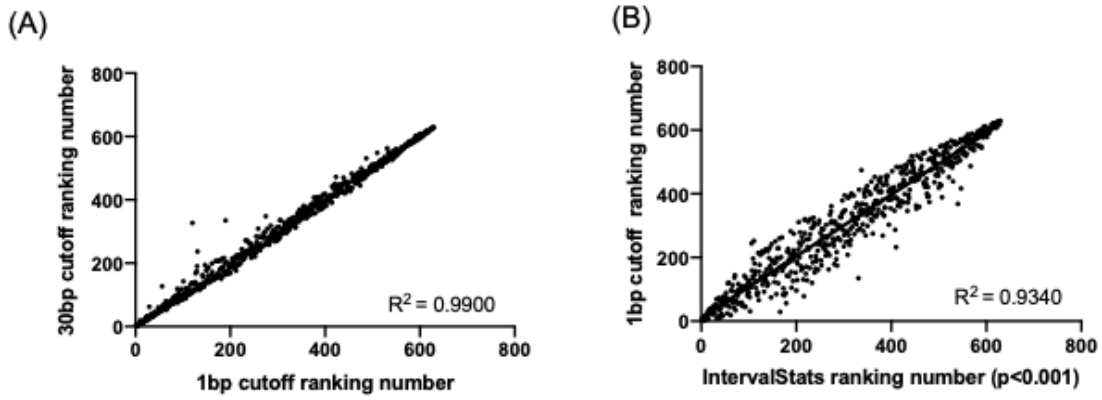

Figure 2.2. Overlapping results comparisons between different approaches or criterion. (A) Consistent results between 30 bp cutoff and 1 bp cutoff. (B) Similar results between 30 bp cutoff based analysis and IntervalStats calculation. Pearson correlation coefficie coefficients were calculated.

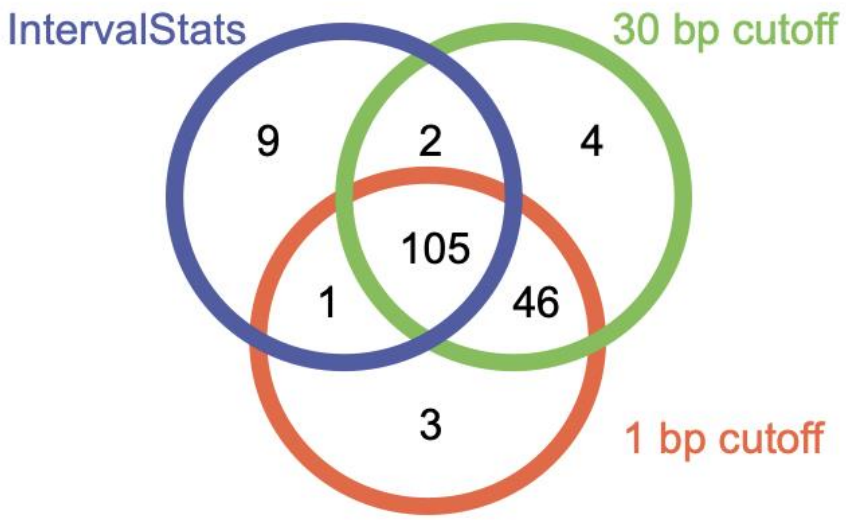

Figure 2.3. A Venn diagram showing that three different overlapping analysis methods yield very similar results. Only those overlaps with $p < 0.0001$ are considered.

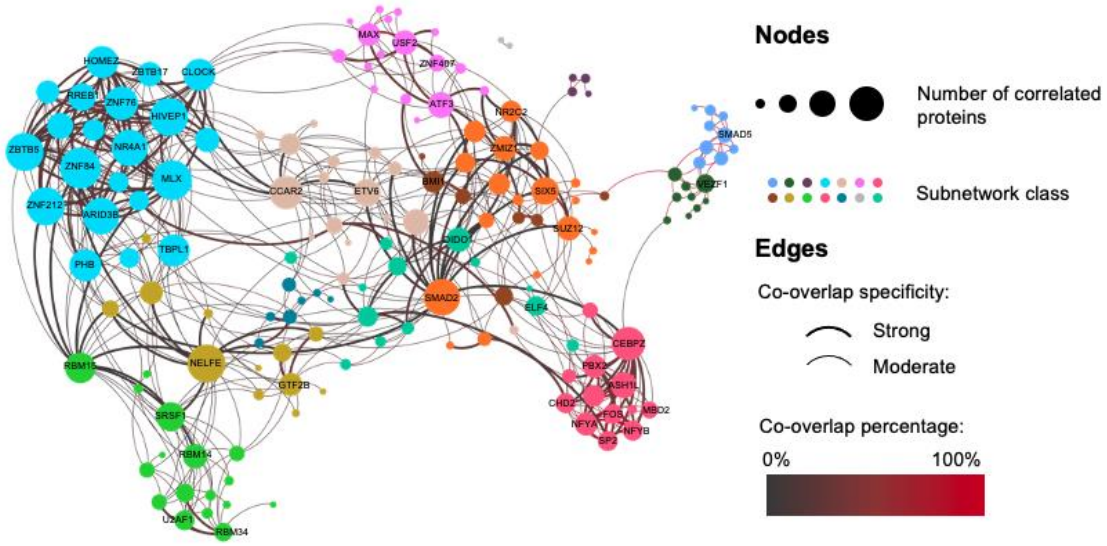

Figure 2.4. Network representations of high confidence co-overlapping upon BG4 binding sites. Each node represents individual potential G4 interaction proteins and colors indicate subnetworks partitioned by algorithm. Edge weight represents co-overlapping specificity while color indicate the co-overlapping percentage. Co-overlapping percentage were calculated based on Jaccard similarity and co-overlapping specificity was determined by identifying outliers.

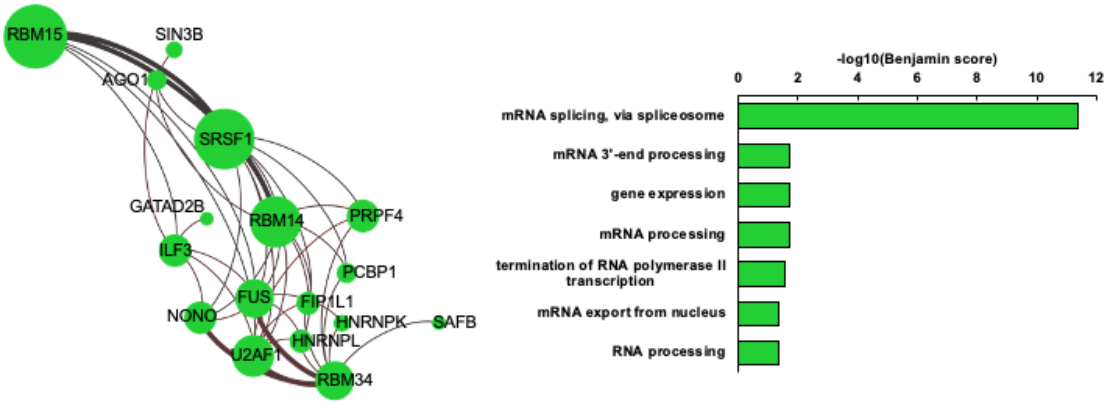

Figure 2.5. Protein cluster with enriched Gene Ontology biological process in mRNA splicing (GO analysis conduced by DAVID and thresholded with adjusted $p < 0.05$.

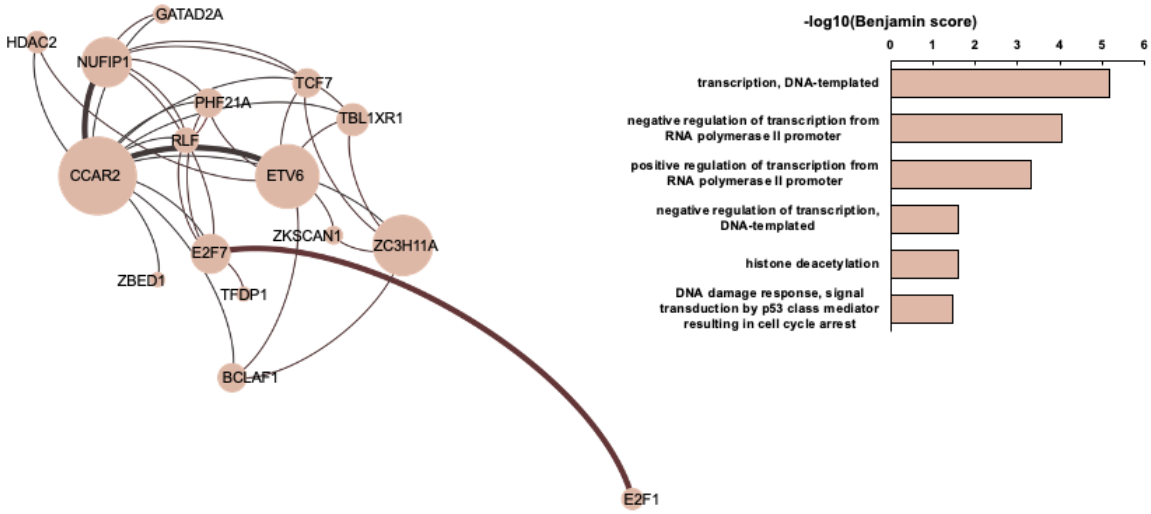

Figure 2.6. Protein cluster with enriched Gene Ontology biological process in transcription regulation (GO analysis conducted by DAVID and thresholded with adjusted $p < 0.05$.

Accession	Target	Cell line	Overlapping Percentage
ENCSR000EWA	H3K4me3	K562	37.16%
ENCSR668LDD	H3K4me3	K562	35.15%
ENCSR000DWD	H3K4me3	K562	34.36%
ENCSR000DUF	H3K4me3	HepG2	28.28%
ENCSR000AKV	H3K9ac	K562	26.05%
ENCSR575RRX	H3K4me3	HepG2	25.93%
ENCSR000AKU	H3K4me3	K562	24.52%
ENCSR000AMD	H3K9ac	HepG2	24.23%
ENCSR000AKP	H3K27ac	K562	22.47%
ENCSR000AMO	H3K27ac	HepG2	21.12%
ENCSR000AMP	H3K4me3	HepG2	18.92%
ENCSR000AOK	H2AFZ	HepG2	17.65%
ENCSR000AKT	H3K4me2	K562	17.63%
ENCSR000EVZ	H3K9ac	K562	12.97%
ENCSR000AMC	H3K4me2	HepG2	12.56%
ENCSR000APC	H2AFZ	K562	12.30%
ENCSR000EWC	H3K4me1	K562	11.24%
ENCSR000APD	H3K79me2	K562	9.50%
ENCSR000AKS	H3K4me1	K562	9.27%
ENCSR000AOM	H3K79me2	HepG2	7.13%
ENCSR000APV	H3K4me1	HepG2	6.45%
ENCSR000AKW	H3K9me1	K562	6.44%
ENCSR000AMQ	H4K20me1	HepG2	5.30%
ENCSR000AKX	H4K20me1	K562	5.19%
ENCSR000AOL	H3K27me3	HepG2	4.03%
ENCSR000AKR	H3K36me3	K562	2.28%
ENCSR000DWB	H3K36me3	K562	2.16%
ENCSR000APE	H3K9me3	K562	1.35%
ENCSR000AMB	H3K36me3	HepG2	0.94%
ENCSR000DUD	H3K36me3	HepG2	0.91%
ENCSR000DUE	H3K27me3	HepG2	0.45%
ENCSR000AKQ	H3K27me3	K562	0.41%
ENCSR000EWB	H3K27me3	K562	0.36%
ENCSR000ATD	H3K9me3	HepG2	0.07%

Table 2.1. Overlapping analysis from ENCODE histone ChIP-seq datasets with BG4 ChIP-seq in K562 and HepG2 cells.

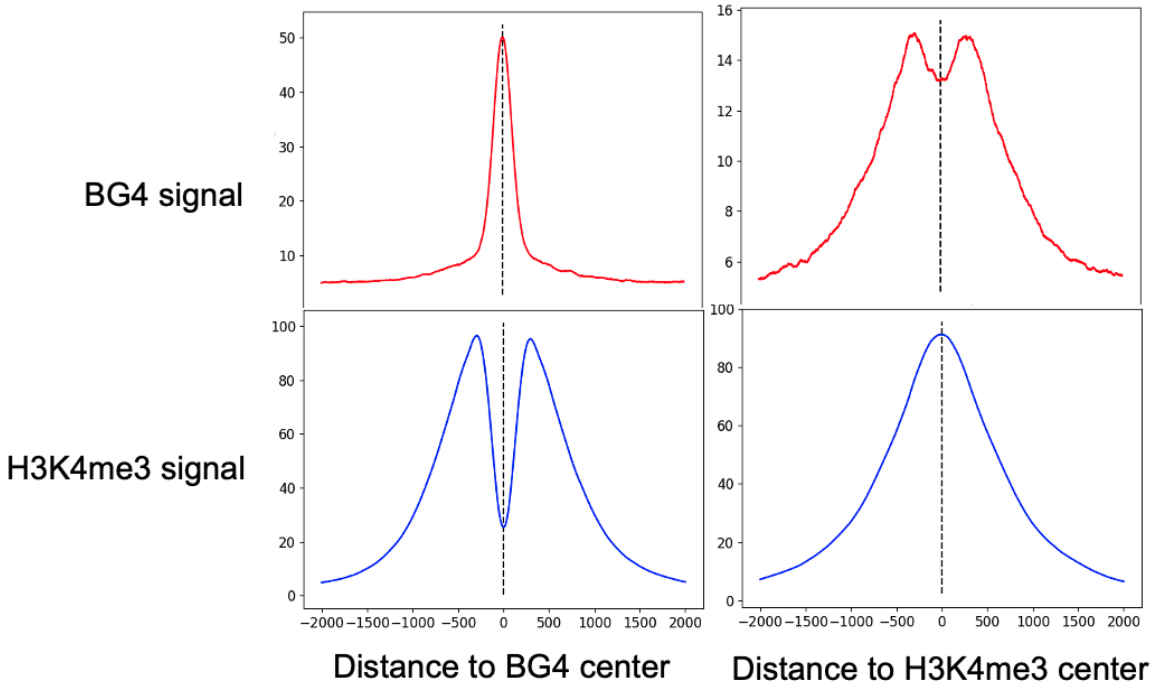

Figure 2.7. Representative enrichment profiles of H3K4me3 and BG4 with respect to the corresponding peak centers.

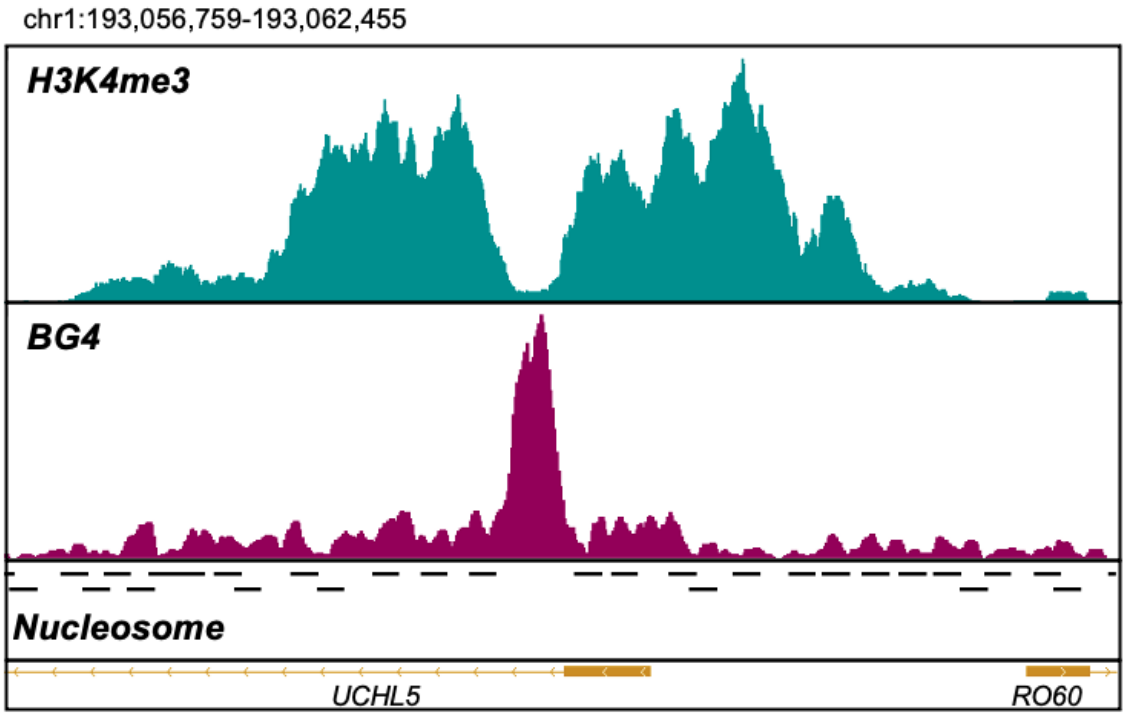

Figure 2.8. Representative IGV plot showing signal track of H3K4me3 and BG4 ChIP-seq, nucleosome position from MNase-seq and gene annotation.

Experiment accession	Cell line	Experiment target	Hit	Total	Percentage
ENCSR529JYA	HepG2	HCFC1	2937	6429	45.68%
ENCSR000EFN	K562	HCFC1	4940	13052	37.85%
ENCSR000AQI	K562	RBBP5	8670	24754	35.02%
ENCSR000EWI	K562	SETDB1	1056	4667	22.63%
ENCSR193ADW	HepG2	SETDB1	302	2484	12.16%
ENCSR642VZY	K562	KDM4B	3390	7999	42.38%
ENCSR872ZHM	HepG2	KDM5A	6056	16196	37.39%
ENCSR000AQA	K562	KDM5B	8141	22368	36.40%
ENCSR620MHD	K562	KDM2B	1665	5205	31.99%
ENCSR280SCF	HepG2	KDM4B	3817	12140	31.44%
ENCSR387JKT	HepG2	KDM3A	5811	18746	31.00%
ENCSR227MRE	HepG2	KDM5B	8364	28624	29.22%
ENCSR754KCC	HepG2	KDM2A	9964	40061	24.87%
ENCSR031ING	HepG2	KDM6A	2967	13996	21.20%
ENCSR087NSR	HepG2	KDM1A	2296	14315	16.04%
ENCSR115BLD	HepG2	KDM1A	7194	51852	13.87%
ENCSR000ATX	K562	KDM1A	1881	30522	6.16%
ENCSR908CMW	K562	KDM1A	1829	43556	4.20%
ENCSR360HRA	K562	KDM1A	1043	31135	3.35%

Table 2.2. Overlapping analysis of histone methylation-modifying enzymes ChIP-seq with BG4 ChIP-seq in K562 and HepG2 cells.

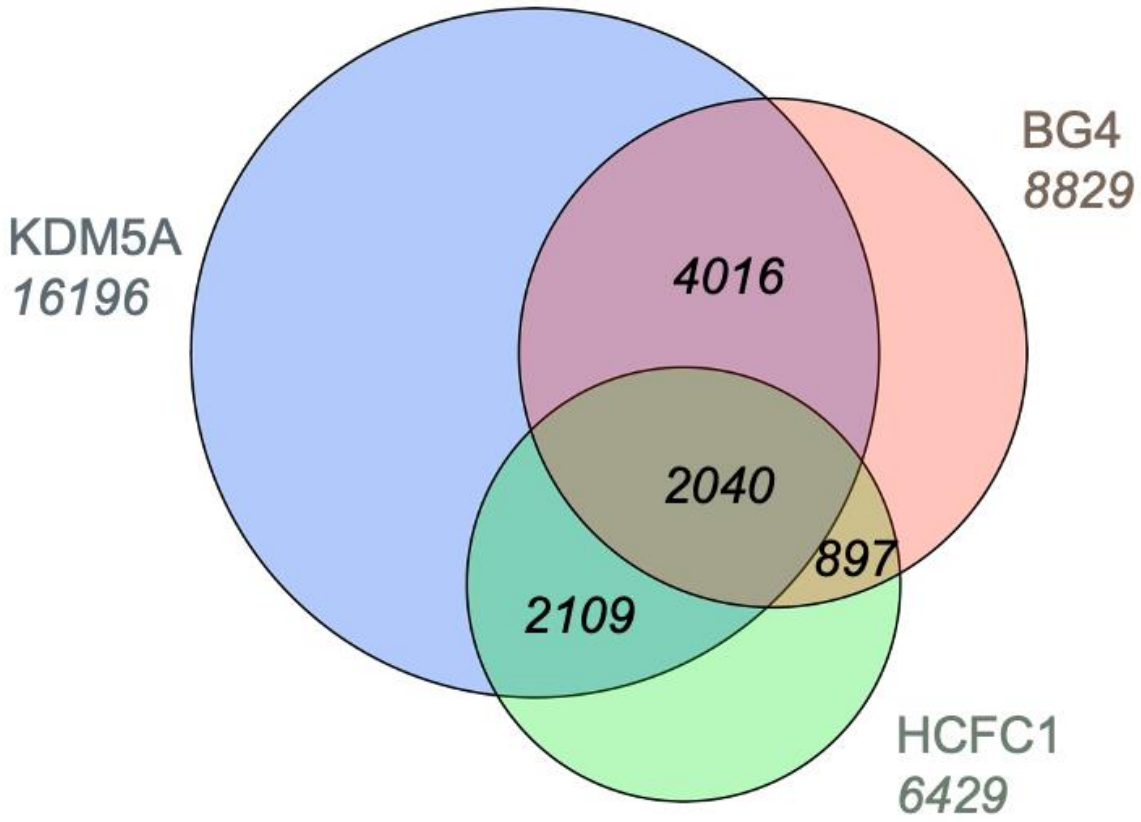

Figure 2.9. A Venn diagram showing overlapping patterns between KDM5B, HCFC1 and BG4 ChIP-seq. Result demonstrated most HCFC1-BG4 overlapping peaks also possess KDM5A binding.

2.5. References

1. Bochman, M.L., Paeschke, K. and Zakian, V.A. (2012) DNA secondary structures: stability and function of G-quadruplex structures. *Nat. Rev. Genet.*, **13**, 770-780.

2. Burge, S., Parkinson, G.N., Hazel, P., Todd, A.K. and Neidle, S. (2006) Quadruplex DNA: sequence, topology and structure. *Nucleic Acids Res.*, **34**, 5402-5415.

3. Mukundan, V.T. and Phan, A.T. (2013) Bulges in G-quadruplexes: broadening the definition of G-quadruplex-forming sequences. *J. Am. Chem. Soc.*, **135**, 5017-5028.

4. Huppert, J.L. and Balasubramanian, S. (2005) Prevalence of quadruplexes in the human genome. *Nucleic Acids Res.*, **33**, 2908-2916.

5. Biffi, G., Tannahill, D., McCafferty, J. and Balasubramanian, S. (2013) Quantitative visualization of DNA G-quadruplex structures in human cells. *Nat. Chem.*, **5**, 182-186.

6. Hansel-Hertsch, R., Beraldi, D., Lensing, S.V., Marsico, G., Zyner, K., Parry, A., Di Antonio, M., Pike, J., Kimura, H., Narita, M. *et al.* (2016) G-quadruplex structures mark human regulatory chromatin. *Nat. Genet.*, **48**, 1267-1272.

7. Hansel-Hertsch, R., Spiegel, J., Marsico, G., Tannahill, D. and Balasubramanian, S. (2018) Genome-wide mapping of endogenous G-quadruplex DNA structures by chromatin immunoprecipitation and high-throughput sequencing. *Nat. Protoc.*, **13**, 551-564.

8. Mao, S.Q., Ghanbarian, A.T., Spiegel, J., Martinez Cuesta, S., Beraldi, D., Di Antonio, M., Marsico, G., Hansel-Hertsch, R., Tannahill, D. and Balasubramanian, S. (2018) DNA G-quadruplex structures mold the DNA methylome. *Nat. Struct. Mol. Biol.*, **25**, 951-957.

9. Spiegel, J., Cuesta, S.M., Adhikari, S., Hansel-Hertsch, R., Tannahill, D. and Balasubramanian, S. (2021) G-quadruplexes are transcription factor binding hubs in human chromatin. *Genome Biol.*, **22**, 117.

10. Bannister, A.J. and Kouzarides, T. (2011) Regulation of chromatin by histone modifications. *Cell Res.*, **21**, 381-395.

11. Xiao, B., Jing, C., Wilson, J.R., Walker, P.A., Vasisht, N., Kelly, G., Howell, S., Taylor, I.A., Blackburn, G.M. and Gamblin, S.J. (2003) Structure and catalytic mechanism of the human histone methyltransferase SET7/9. *Nature*, **421**, 652-656.

12. Tamaru, H., Zhang, X., McMillen, D., Singh, P.B., Nakayama, J., Grewal, S.I., Allis, C.D., Cheng, X. and Selker, E.U. (2003) Trimethylated lysine 9 of histone H3 is a mark for DNA methylation in Neurospora crassa. *Nat. Genet.*, **34**, 75-79.

13. D'Oto, A., Tian, Q.W., Davidoff, A.M. and Yang, J. (2016) Histone demethylases and their roles in cancer epigenetics. *J Med Oncol Ther*, **1**, 34-40.

14. Greer, E.L. and Shi, Y. (2012) Histone methylation: a dynamic mark in health, disease and inheritance. *Nat. Rev. Genet.*, **13**, 343-357.

15. Davis, C.A., Hitz, B.C., Sloan, C.A., Chan, E.T., Davidson, J.M., Gabdank, I., Hilton, J.A., Jain, K., Baymuradov, U.K., Narayanan, A.K. *et al.* (2018) The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res.*, **46**, D794-D801.

16. Chikina, M.D. and Troyanskaya, O.G. (2012) An effective statistical evaluation of ChIPseq dataset similarity. *Bioinformatics*, **28**, 607-613.

17. Griffon, A., Barbier, Q., Dalino, J., van Helden, J., Spicuglia, S. and Ballester, B. (2015) Integrative analysis of public ChIP-seq experiments reveals a complex multi-cell regulatory landscape. *Nucleic Acids Res.*, **43**, e27.

18. Bastian, M., Heymann, S. and Jacomy, M. (2009), *Third international AAAI conference on weblogs and social media*.

19. Blondel, V.D., Guillaume, J.L., Lambiotte, R. and Lefebvre, E. (2008) Fast unfolding of communities in large networks. *Journal of Statistical Mechanics-Theory and Experiment*, **2008**, P10008.

20. Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H. and 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**, 576-589.

21. Huang da, W., Sherman, B.T. and Lempicki, R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44-57.

22. Sherman, B.T., Hao, M., Qiu, J., Jiao, X., Baseler, M.W., Lane, H.C., Imamichi, T. and Chang, W. (2022) DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Res.*, **50**, W216-W221.

23. Ross-Innes, C.S., Stark, R., Teschendorff, A.E., Holmes, K.A., Ali, H.R., Dunning, M.J., Brown, G.D., Gojis, O., Ellis, I.O., Green, A.R. *et al.* (2012) Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature*, **481**, 389-393.

24. Pohl, A. and Beato, M. (2014) bwtool: a tool for bigWig files. *Bioinformatics*, **30**, 1618-1619.

25. Simko, E.A.J., Liu, H., Zhang, T., Velasquez, A., Teli, S., Haeusler, A.R. and Wang, J. (2020) G-quadruplexes offer a conserved structural motif for NONO recruitment to NEAT1 architectural lncRNA. *Nucleic Acids Res.*, **48**, 7421-7438.

26. Takahama, K., Takada, A., Tada, S., Shimizu, M., Sayama, K., Kurokawa, R. and Oyoshi, T. (2013) Regulation of telomere length by G-quadruplex telomere DNA- and TERRA-binding protein TLS/FUS. *Chem Biol*, **20**, 341-350.

27. Lai, A.Y. and Wade, P.A. (2011) Cancer biology and NuRD: a multifaceted chromatin remodelling complex. *Nat. Rev. Cancer*, **11**, 588-596.

28. Zheng, D., Trynda, J., Sun, Z. and Li, Z. (2019) NUCLIZE for quantifying epigenome: generating histone modification data at single-nucleosome resolution using genuine nucleosome positions. *BMC Genomics*, **20**, 541.

29. Schultz, D.C., Ayyanathan, K., Negorev, D., Maul, G.G. and Rauscher, F.J. (2002) SETDB1: a novel KAP-1-associated histone H3, lysine 9-specific methyltransferase that contributes to HP1-mediated silencing of euchromatic genes by KRAB zinc-finger proteins. *Genes Dev.*, **16**, 919-932.

30. Karanth, A.V., Maniswami, R.R., Prashanth, S., Govindaraj, H., Padmavathy, R., Jegatheesan, S.K., Mullangi, R. and Rajagopal, S. (2017) Emerging role of SETDB1 as a therapeutic target. *Expert Opin. Ther. Targets*, **21**, 319-331.

31. Robinson, J., Raguseo, F., Nuccio, S.P., Liano, D. and Di Antonio, M. (2021) DNA G-quadruplex structures: more than simple roadblocks to transcription? *Nucleic Acids Res.*, **49**, 8419-8431.

32. Mukherjee, A.K., Sharma, S., Bagri, S., Kutum, R., Kumar, P., Hussain, A., Singh, P., Saha, D., Kar, A., Dash, D. *et al.* (2019) Telomere repeat-binding factor 2 binds extensively to extra-telomeric G-quadruplexes and regulates the epigenetic status of several gene promoters. *J. Biol. Chem.*, **294**, 17709-17722.

33. Krasnov, A.N., Mazina, M.Y., Nikolenko, J.V. and Vorobyeva, N.E. (2016) On the way of revealing coactivator complexes cross-talk during transcriptional activation. *Cell Biosci.*, **6**, 15.

34. Furey, T.S. (2012) ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. *Nat. Rev. Genet.*, **13**, 840-852.

Chapter 3: G-quadruplex DNA Contributes to RNA Polymerase II-mediated 3D Chromatin Architecture

3.1. Introduction

The DNA guanine quadruplexes (G4) are four-stranded secondary structures that form in guanine-rich regions of DNA (1,2). G4 structures arise from stacking of at least three layers of G tetrads, where four Gs are held together through hydrogen bonding and a monovalent cation, in the order of $K^+ > Na^+ > Li^+$ (3). The development of a G4 structure-specific antibody (BG4) enabled probing for the presence of G4 structures in chromatin (4), where BG4 ChIP-seq results unveiled the enrichment of G4s in regulatory regions of the genome and the presence of distinct G4 landscapes in different cell lines (5,6).

A number of studies revealed the important roles of G4 in various cellular processes, including telomere maintenance (7,8), DNA replication (9), DNA damage (10), and transcription regulation (11-13). Integrative analysis of ChIP-seq data of transcription factors and G4 structures underscored G4s as binding hubs for transcription factors in cells (14). In addition, multiple proteins were shown to bind directly to G4 structures *in vitro* (13,15-19).

Another important element of gene regulation is 3D chromatin architecture, where the nucleus of mammalian cells is highly compacted and organized in a hierarchical fashion ranging from A-B compartment to enhancer-promoter (E-P) contacts (20). Extensive studies have been conducted to investigate the functions of E-P interactions in transcription regulation and their potential contributions to disease development (21-24). In this vein, multiple methods have been developed to profile the detailed 3D organizations of the human genome, including 3C, 4C, 5C,

HiC and HiChIP-seq/ChIA-PET/PLAC-seq (25). Among these methods, HiChIP-seq/ChIA-PET/PLAC-seq can detect specific protein-mediated long-range DNA interactions and provide important information about how individual proteins modulate high-order genome organization. Bioinformatic analysis showed the enrichment of G4 structures in topologically-associating domain (TAD) boundaries and in E-P interactions, indicating the role of G4s in high-order chromatin organization (26). Moreover, a recent study revealed that Yin-Yang 1 (YY1), which is known to dimerize and enable E-P interactions (27), is able to interact with G4 DNA at high affinity (13). Further HiChIP-Seq assay substantiated a YY1-mediated, G4-dependent DNA looping (13). Although multiple studies suggest a role of G4s in distal gene regulation, not much is known about the detailed mechanisms through which G4 structures modulate long-range DNA interactions in cells.

Here, by conducting an overlapping-based analysis of ChIA-PET data in publicly available Encyclopedia of DNA Elements (ENCODE) database (28), we identified a strong correlation between G4 structures in chromatin and RNAPII-mediated long-range DNA interactions. We also observed that treatment of cells with pyridostatin (PDS), a small-molecule G4-binding ligand, led to more marked decreases in RNAPII-mediated DNA looping at sites with G4 structures than those without. In addition, genome-wide association analysis between ChIA-PET/HiChIP-seq and RNA-seq data provided a comprehensive understanding about transcription regulation mediated by the interplay of G4 structure and RNAPII-mediated DNA looping. Moreover, we showed that enhancer G4s modulate the expression of *AKR1C* family genes in HepG2 cells.

3.2. Materials and Methods

Cell lines

HepG2 human hepatocellular carcinoma cells were cultured in Dulbecco's modified Eagle's medium (DMEM, life Technologies) containing 10% fetal bovine serum (Invitrogen) and 1% penicillin and streptomycin (Invitrogen). K562 human chronic myelogenous leukemia cells were cultured in RPMI 1640 medium (Life Technologies) containing 10% fetal bovine serum (Invitrogen) and 1% penicillin and streptomycin (Invitrogen). The cells were maintained at 37°C in an incubator containing 5% CO_2 and the cells were tested to be free of mycoplasma contamination using e-Myco PCR Detection Kits (Bulldog Bio).

Cell viability assay

Cell viability was examined using Cell Counting Kit-8 (CCK8, Dojindo) according to the manufacturer's recommended procedures, where a 100-μL suspension containing 5000 HepG2 cells was plated in each well of a 96-well plate one day prior to treatment. Ten μL of the indicated concentrations of PDS were added to each well, and the cells were incubated for 24 h. After the incubation, 10 μL of CCK-8 solution was added to each well, and absorbance at 460 nm was recorded 3 h later with a BioTek Synergy H1 microplate reader (Agilent Technologies).

Bioinformatic analysis

ChIA-PET datasets were retrieved through the ENCODE portal under assay title "ChIA-PET", target of assay "POLR2A" and biosample "K562" or "HepG2". POLR2A ChIP-seq datasets were also downloaded from the ENCODE portal under assay title "ChIP-seq". Bedpe files in GRCh38 assembly were downloaded for overlapping analysis and contact matrix hic files were used for visualization. BG4-ChIP-seq raw data for HepG2 and K562 cells were obtained from Sequence Read Archive (SRA) with the accession number of PRNJ60617 (14). BG4-ChIP-seq data were processed following previously published procedures in GRCh38 assembly (14). Overlapping percentages between RNAPII-linked long-range DNA interactions and G4s were calculated using bedtools pairToBed command with different -type parameters (29). One to multiple overlaps were

combined accordingly. Bedpe files were split into two anchors and deduplicated to produce loop anchors. POLR2A ChIP-seq narrowPeak files were overlapped with loop anchors to obtain POLR2A binding sites with or without long-range interactions. Overlap between peaks was calculated using bedtools intersect command. Monte Carlo simulation was conducted by randomly shuffling peak file in target regions using bedtools shuffle command.

HiChIP and data analysis

HiChIP was performed as previously described (30). Ten million HepG2 cells were mock-treated (with sterilized water) or treated with 20 μM PDS for 24 h before crosslinked with a freshly prepared 1% formaldehyde solution at room temperature for 10 min. After quenching with glycine at a final concentration of 125 mM for 10 min, the cells were washed several times with PBS buffer and subsequently incubated in HiChIP lysis buffer (10 mM Tris-HCl, pH 8.0, 10 mM NaCl, 0.2% NP-40, and freshly added protease inhibitor cocktail) at 4°C for 2 h with rotation. After washing once with cold HiChIP lysis buffer and centrifugation, the cell pellet was resuspended in 0.5% SDS (100 μL) and incubated at 62°C for 10 min. SDS was later quenched by adding 25 μL of freshly prepared 10% Triton X-100 and 135 μL water. After incubation at 37°C for 15 min, the resulting chromatin was restriction-digested by adding 25 μL 10x rCutsmart buffer (NEB) and 100 units MboI (NEB). Chromatin was digested overnight at 37°C with shaking at 900 rpm. MboI was inactivated by incubation at 62°C for 20 min and then cooling to room temperature. To the mixture were subsequently added 15 nmol each of dCTP, dGTP, dTTP (NEB), biotin-14-dATP (Jena Bioscience) and 40 U Klenow fragment (NEB) in a total volume of 300 μL to perform nucleotide fill-in and biotin labelling. Following incubation at 37°C with shaking at 900 rpm for 1 h, a DNA ligase master mix, which contained 664 μL water, 120 μL 10 x T4 ligase buffer (NEB), 10% Triton X-100, 6 μL 20 mg/μL BSA, and 10 μL T4 DNA ligase (NEB), was added to the reaction mixture, and the mixture was incubated at room temperature for 6 h. The chromatin was collected by

centrifugation and sonicated into 300-500 bp DNA fragments in RIPA buffer (10 mM Tris-HCl, pH 8.0, 140 mM NaCl, 1 mM EDTA, 1% Triton X-100, 0.1% SDS, 0.1% sodium deoxycholate) and then incubated with 10 μg POLR2A antibody (Thermo Fisher Scientific) at 4°C overnight. Antibody-bound chromatin was captured by 50 μL Protein-A/G magnetic beads (Thermo Fisher Scientific) pre-blocked with PBS/BSA (5 mg/mL BSA in 1xPBS). The beads were subsequently washed with a low-salt RIPA buffer (10 mM Tris-HCl, pH 8.0, 140 mM NaCl, 1 mM EDTA, 1% Triton X-100, 0.1% SDS, 0.1% sodium deoxycholate) three times, a high-salt RIPA buffer (10 mM Tris-HCl, pH 8.0, 300 mM NaCl, 1 mM EDTA, 1% Triton X-100, 0.1% SDS, 0.1% sodium deoxycholate with proteinase inhibitor cocktail) three times, a LiCl washing buffer (10 mM Tris-HCl, pH 8.0, 150 mM LiCl, 1 mM EDTA, 0.5% NP-40, 0.1% sodium deoxycholate) three times and a TE buffer (10 mM Tris-HCl, pH 8.0, 0.1 mM EDTA) twice. DNA was purified by DNA Clean & Concentrator-5 (Zymo Research), and subsequently quantified using Qubit (Thermo Fisher Scientific). Biotin-labeled DNA was enriched using the Dynabeads MyOne streptavidin C1 (Thermo Fisher Scientific). For each sample, 20 μL streptavidin C1 beads were washed twice with 400 μL Tween wash buffer (5 mM Tris-HCl, pH 7.5, 0.5 mM EDTA, 1 M NaCl, 0.05% Tween 20) and resuspended in 40 μL 2x binding buffer (10 mM Tris-HCl, pH 7.5, 1.0 mM EDTA, and 2.0 M NaCl) and incubated with 40 μL of the above-mentioned DNA isolated from ChIP procedures at room temperature for 15 min. The DNA-bound beads were washed twice with 500 μL Tween wash buffer, collected and resuspended in 25 μL TE buffer. Sequencing libraries were constructed using NEBNext Ultra DNA Library Prep Kit for Illumina (NEB) following the manufacturer's instructions. The library was paired-end sequenced (2x100bp) on a MGI 2000 platform (BGI). Two biological replicates were performed for each condition.

Paired-end reads were processed using HiC-Pro with default parameter (version 3.1.0) (31). Fastq files were aligned to GRCh38 assembly, deduplicated, and assigned to MboI restriction

fragments. After filtration for valid interactions, interaction matrices were constructed and visualized using HiCExplorer (32).

RNA-seq and data analysis

Total RNA was extracted from cells using Omega Total RNA Kit I accordingly to the manufacturer's recommended procedures. Poly(A) RNA enrichment was conducted using NEBNext Poly(A) mRNA Magnetic Isolation Module (NEB), and the sequencing library was constructed by using the NEBNext Ultra II RNA Library Prep Kit (NEB) following the manufacturer's instructions. The resulting library was subjected to sequencing analysis on a MGI 2000 platform. The sequencing reads were aligned to GRCh38 assembly using STAR (v.2.7.0) with default parameters (33). Transcript quantification was conducted using featureCounts (v.2.0.3) (34). Differential gene expression analysis was performed with DESeq2 (v.1.36.0) (35).

Real-time quantitative PCR (RT-qPCR)

Total RNA was extracted using Omega Total RNA Kit I (Omega) following the vendor's recommended procedures and quantified. Approximately 2 μg total RNA was immediately reverse-transcribed using 200 units of M-MLV reverse transcriptase (Promega) with 1.0 μg oligo$(dT)_{20}$ primer according to manufacturer's recommended procedures. RT-qPCR experiments were performed using Luna Universal qPCR Master Mix (NEB) on a CFX96 RT-qPCR detection system (Bio-rad), by following the manufacturer's recommended protocol. Standard curves of each gene amplification product were obtained. Correlation coefficients for the standard curves were confirmed to be at least 0.99, and the amplification efficiencies were verified to be within 90%-110%. Relative quantifications of the genes of interest were conducted based on standard curves and normalized to *GAPDH*. Primers used in RT-qPCR are listed in.

Chromosome conformation capture-qPCR (3C-qPCR)

3C-qPCR was performed as previously described (36) with some modifications. Briefly, 10 million HepG2 cells (mock-treated or treated with 20 μM PDS for 24 h; DMSO-treated or treated with 1 μM JQ1 for 24 h) were crosslinked in freshly prepared 1% formaldehyde in PBS buffer at room temperature for 10 min and then quenched by incubating with 125 mM glycine at room temperature for 10 min. The cells were harvested, and the cell pellet was suspended in 1 mL cold lysis buffer (10 mM Tris-HCl, pH 7.5, 10 mM NaCl, 0.2% NP-40 with freshly added protease inhibitor cocktail) and incubated with rotation at 4°C for 3 h. After centrifugation at 400 g at 4°C for 5 min, the resulting nuclear pellet was resuspended in 0.5 mL of 1.2x restriction enzyme buffer (60 μL 10x rCutsmart buffer and 440 μL H_2O) and transferred to a 37°C thermomixer. To the reaction mixture was then added 7.5 μL of 20% (w/v) SDS, and the mixture was incubated at 37°C with shaking at 900 rpm for 1 h. After quenching the SDS with 50 μL 20% (v/v) Triton X-100 at 37°C for 1 h, the reaction mixture was digested overnight with 400 U EcoRI-HF (NEB) at 37°C with shaking at 900 rpm. Another round of digestion was performed by adding a new aliquot of EcoRI-HF (400 U) to the reaction mixture the next day, and the mixture was incubated at 37°C with shaking at 900 rpm for 2 h. The restriction enzyme was subsequently deactivated by addition of 40 μL 20% (w/v) SDS and incubation at 65°C for 25 min. The reaction mixture was then diluted with 700 μL 10x T4 ligase buffer, 5.425 mL ddH_2O and 375 μL 20% (w/v) Triton X-100 and incubated at 37°C with gentle shaking for 1 h. To the resulting mixture was added 2000 U T4 ligase (NEB), and the mixture was incubated at 16°C overnight. The sample was then treated with 300 μg proteinase K and the crosslink was reversed by heating at 65°C overnight. RNA was removed by incubating with 300 μg RNase A at 37°C for 1 h. DNA was purified by phenol-chloroform extraction. Real-time PCR quantifications of ligation products were performed using Luna Universal qPCR Master Mix (NEB) on a CFX96 RT-qPCR detection system (Bio-rad) following

the manufacturer's recommended protocol. A digested and re-ligated bacterial artificial chromosome (BAC CH17-30P14), covering the genomic regions of interest, was used as a control template. Primers were designed to be in the same direction and as close to the EcoRI restriction sites as possible. A constant primer and a test primer were used in each qPCR reaction. Standard curves of ligation products were constructed using serial dilution of control template. The 3C-qPCR data were normalized to a control interaction localized in the *ERCC3* gene.

Chromatin immunoprecipitation-qPCR (ChIP-qPCR)

ChIP experiments were conducted as preciously described with a few modifications (37). Briefly, 10 million HepG2 cells were crosslinked in freshly prepared 1% formaldehyde in PBS buffer at room temperature for 10 min and then quenched by incubating with 125 mM glycine at room temperature for 10 min. The cells were harvested, and the cell pellet was suspended in 1 mL cold lysis buffer (10 mM Tris-HCl, pH 7.5, 10 mM NaCl, 0.2% NP-40 with freshly added protease inhibitor cocktail) and incubated with rotation at 4°C for 1 h. After centrifugation at 400 g at 4°C for 5 min, the resulting nuclear pellet was resuspended in RIPA buffer (10 mM Tris-HCl, pH 8.0, 140 mM NaCl, 1 mM EDTA, 1% Triton X-100, 0.1% SDS, 0.1% sodium deoxycholate) and incubated with rotation at 4°C for 30 min. Chromatin was sheared using a QSONICA sonicator Q125 at 4°C for 10 min (10 sec on / 10 sec off pulse) with a 42% amplitude. The resulting mixture was centrifuged at 13,200 g at 4°C for 15 min. The supernatant was incubated with 5 μg POLR2A antibody (Thermo Fisher Scientific) at 4°C overnight. Antibody-bound chromatin was captured by 50 μL Protein-A/G magnetic beads (Thermo Fisher Scientific). The beads were subsequently washed with a low-salt RIPA buffer (10 mM Tris-HCl, pH 8.0, 140 mM NaCl, 1 mM EDTA, 1% Triton X-100, 0.1% SDS, 0.1% sodium deoxycholate) three times, a high-salt RIPA buffer (10 mM Tris-HCl, pH 8.0, 300 mM NaCl, 1 mM EDTA, 1% Triton X-100, 0.1% SDS, 0.1% sodium deoxycholate with proteinase inhibitor cocktail) once, a LiCl washing buffer (10 mM Tris-HCl, pH

8.0, 150 mM LiCl, 1 mM EDTA, 0.5% NP-40, 0.1% sodium deoxycholate) once and a TE buffer (10 mM Tris-HCl, pH 8.0, 0.1 mM EDTA) once. DNA was purified by DNA Clean & Concentrator-5 (Zymo Research). Quantitative PCR was conducted using Luna Universal qPCR Master Mix (NEB) on a CFX96 RT-qPCR detection system (Bio-rad) following the manufacturer's recommended protocol.

3.3. Results

Overlapping analysis revealed a correlation between RNAPII-mediated long-range DNA interaction and G4 structures

To investigate the correlation between RNAPII-dependent 3D genome organization and DNA G4, we assessed the co-occupancy of endogenous G4 structure loci with the two anchors of RNAPII-linked DNA loops identified from ChIA-PET analysis (38). To this end, we performed overlapping analysis using POLR2A ChIA-PET data retrieved from the ENCODE database and BG4-ChIP-seq results obtained for the same cell lines, i.e., HepG2, K562 and HEK293T cells (6,14,28,39). Our results showed that large percentages of DNA loops (141010/220992, 63.8% in HepG2 cells; 59729/186714, 32.0% in K562 cells; and 96902/174673, 55.5% in HEK293T cells) carried at least one G4 structure in the two anchors (Figure 3.1A). When compared to CTCF ChIA-PET data, we observed a higher overlapping percentage of RNAPII-mediated long-range DNA interactions with endogenous G4 sites, indicating active engagement of DNA G4 structures in transcription. Additionally, the majority of endogenous G4 sites (19979/28382, 70.4% in HepG2 cells; 12676/19238, 65.9% in K562 cells; and 12438/19965, 62.3% in HEK293T cells) are associated with RNAPII-linked long-range DNA interactions (Figure 3.1B**)**, supporting a positive correlation between G4 structure and RNAPII-mediated DNA looping.

We also analyzed the RNAPII-linked DNA loops in HepG2 cells and found that they encompass not only promoter-enhancer interactions (24.1%), but also promoter-promoter (8.8%), enhancer-enhancer interactions (24.4%), and those not involving annotated promoters or enhancers (42.7%). Similar findings were made for those RNAPII-associated DNA loops with at least one anchor containing G4 structure, though with slightly higher percentages promoter-promoter (12.7%) and promoter-enhancer (31.5%) interactions.

Because there is a significant enrichment of GC content in promoter region of human genes (40) , it is important to examine if such a positive correlation arises from a bias in sequence context or open chromatin. To explore this possibility, we randomly chose the same number of peaks in control regions that are comprised of GC-rich sequences and exhibit G4-forming potential *in vitro*, i.e., those regions with observed G-quadruplex sequences (OQS) (41). In both K562 and HepG2 cells, RNAPII-linked DNA loops display a significantly weaker co-occurrence with randomly picked regions with OQS ($p < 0.01$, Monte Carlo Simulation) than with those loci enriched with G4 structures. We also conducted a similar analysis for open chromatin as reflected by DNase I hypersensitive sites. The results showed that the RNAPII-linked DNA loops displayed mean overlapping percentages of 35.6% and 11.4% with DNase I hypersensitive sites in HepG2 and K562 cells, respectively, which are substantially lower than with those sites harboring G4 structures ($p < 0.01$. Monte Carlo Simulation). These results suggest that the enrichment of RNAPII-linked DNA loops at endogenous G4 structure loci is not simply due to the primary sequence of DNA elements at those sites (i.e., being GC-rich) or the association of those loci with open chromatin, but rather attributed to the formation of G4 structures at these sites.

G4s were proposed to be binding hubs for transcription factors to promote active transcription (14). As RNAPII-binding sites are directly associated with transcription activity, we assessed the G4 overlapping at those RNAPII-binding sites, as revealed by ChIP-seq data, with

long-range DNA interactions and those without. Our results showed a much higher G4 percentage at those RNAPII-binding loci that are involved with DNA looping (Figure 3.1C), underscoring the role of G4 in RNAPII-associated long-range DNA interactions.

To further investigate the relationship between DNA looping and G4, we calculated the DNA interaction PET numbers measured by ChIA-PET assay on the basis of G4 overlapping status. The results from HepG2, HEK293T and K562 cells showed significantly higher interaction frequencies when one anchor of DNA loops overlapped with G4 sites in chromatin compared to those loops not associated with cellular G4s. Additionally, more pronounced interactions were detected in cases where both loop anchors carry G4 structures (Figure 3.1D). These results indicate an active participation of G4s in RNAPII-mediated long-range DNA interactions.

G4 landscapes are distinct in different cells (6), and so are 3D genome organizations (42). Thus, we queried, using Diffbind algorithm (43), differential G4 sites in HepG2 and K562 cells based on statistically significant differences in BG4 ChIP signal. A total of 18,991 differential G4 structure sites were retrieved, including 9726 and 9265 in HepG2 and K562 cells, respectively. Likewise, a total of 10956 cell type-specific DNA loops were called from ChIA-PET datasets for the two cell lines. We next compared the overlapping pattern between differential G4s and long-range DNA interactions (44). Those DNA loops preferentially detected in HepG2 cells are more likely to overlap with G4 structures specifically detected in HepG2 cells than those detected uniquely in K562 cells (4378 vs. 506, total 6598); the same finding was made for DNA loops preferentially detected in K562 cells (2347 vs. 168, total 4387). These data again suggest the involvement of cellular G4 structures in RNAPII-mediated long-range DNA interactions.

PDS preferentially diminished RNAPII-mediated long-range DNA interactions involving G4 structure loci

To explore further the origins of the positive correlation between G4 structures and RNAPII-mediated DNA looping, we conducted POLR2A HiChIP-seq in HepG2 cells with or without PDS treatment. In this context, PDS, a small-molecule G4 ligand that binds specifically to G4 structures, has been widely used for displacing G4-binding proteins from G4 sites in cells (13,45,46). As HiChIP-seq also captures the *in vivo* binding landscape of target proteins, we first examined the effect of PDS on RNAPII binding profiles. A total of 22978 RNAPII peaks were identified in "mock" condition and exhibited a high overlapping percentage (10467/22978, 45.5%) with BG4 ChIP-seq peaks. Following a 24-h treatment with 20 μM PDS, only 7278 RNAPII binding sites were captured and overlapping analysis showed an attenuated co-occurrence of the RNAPII binding loci with endogenous G4 structures (2609/7278, 35.8%) (Figure 3.2A). Likewise, signal intensities of RNAPII are strongly diminished in cells treated with PDS than those without (mock), where the signal ratio of PDS/mock for G4 loci were significantly lower than non-G4 loci (Figure 3.2B), indicating a role of PDS in impairing the recruitment of RNAPII to G4 structure sites in cells. IGV plots of representative regions showed a strong effect of PDS on displacing RNAPII from the promoters of *SLC26A2*, *TIGD6* and *HMGXB3* genes that carry G4 structures (Figure 3.2C). However, RNAPII ChIP signal was augmented in non-G4 regions after PDS treatment, which is consistent with the above-mentioned statistical analysis. In this vein, our cell viability assay results showed that a 24-h exposure with 20 μM did not give rise to any apparent alteration in the viability of HepG2 cells. We next examined the impact of G4 structures on RNAPII-mediated long-range DNA interactions. The results from HiChIP-seq analysis showed that PDS treatment markedly attenuated RNAPII-mediated DNA looping, as shown in the chromosome matrix view (Figure 3.3A). In addition, a pronouncedly decreased number of DNA loops were detected in PDS-treated cells relative to mock-treated cells (66417 vs. 18778), supporting that G4 is crucial for RNAPII-linked long-range DNA contacts. We also observed an attenuated presence

of G4 structure sites in the loop anchors of the detected long-range DNA interactions in PDS-relative to mock-treated HepG2 cells (27.7% vs. 33.8%). Moreover, those DNA loops not perturbed by PDS treatment exhibited a much lower extent of overlap with endogenous G4 loci, indicating that PDS preferentially disrupts G4-mediated long-range DNA interactions.

Genome-wide accumulation analysis of RNAPII-associated DNA loops in a distance range of 5-200 kilobases (kb) showed much weaker RNAPII-linked DNA interactions following PDS treatment (Figure 3.3B). Furthermore, those RNAPII-mediated DNA loops overlapped with endogenous G4 loci displayed a more pronounced diminution in long-range interaction frequency in cells treated with PDS than those lacking overlap with endogenous G4 sites (Figure 3.3C).

Previous studies showed the enrichment of G4 structures at the promoters of oncogenes, including *KRAS* and *MDM2* (47,48). Thus, we examined whether G4 structures play any role in RNAPII-mediated DNA interactions of those oncogenes. Analysis of our HiChIP-seq data revealed multiple G4-containing long-range DNA interactions involving the promoters of *KRAS* and *MDM2* genes, and the disruption of these interactions following PDS treatment (Figure 3.3D). In this vein, our RNAPII ChIP-qPCR experiment revealed enrichments of RNA polymerase II at both the promoters and enhancers of *KRAS* and *MDM2* genes, which are also enriched with G4 structures (Figure 3.3D).

Together, these data furnish evidence to support that RNAPII-mediated long-range DNA interactions are highly associated with G4 structures, and G4-binding ligand could perturb RNAPII-linked and G4-dependent DNA loops, including those involving promoters of oncogenes.

G4-dependent DNA loops regulate gene expression

RNAPII is responsible for mRNA transcription and plays a vital role in gene expression (49). In light of the above results showing that cellular G4s are highly correlated with RNAPII-

mediated DNA loops, we next asked whether G4-dependent long-range DNA interactions modulate gene expression.

As G4 structures are proposed to be correlated with active transcription (5,11,14,19), we first evaluated the expression pattern of genes with respect to the presence of G4 structures (based on BG4 ChIP-seq data) and long-range DNA interactions (based on ChIA-PET data). We divided genes into four groups based on their associations with G4 structures (Figure 3.4A): genes in *Group A* carry G4 structures in their promoters; *Groups B* and *C* genes do not contain G4 structures in their promoters, but form loops with a distal site with (*Group B*) or without (*Group C*) G4 structure; and the remaining genes were classified into *Group D*. Among these four groups of genes, *Group A* exhibits the highest expression level, and *Group B* displays significantly higher expression profile than *Group C*. Those genes without any RNAPII-linked long-range interactions or associated with G4 structures (*Group D*) exhibit the lowest expression profile (Figure 3.4B).

Next, we evaluated the influence of PDS treatment on cellular transcriptome. We found that genes in the aforementioned *Groups A* and *B* exhibited diminished expression after PDS treatment; such diminished expression, however, is much less pronounced for *Group C* genes, and not observed for those in *Group D* (Figure 3.4C). These data underscored that G4 structures not only locally modulate expression of target genes through their promoters, but also distally regulate the transcription of target genes through RNAPII-linked long-range DNA interactions.

G4-dependent DNA loops activate the expression of *AKR1C* family genes

The *AKR1C1-AKR1C3* genes are closely located on chromosome 10 in a region spanning ~ 200 kb. RNAPII ChIA-PET data in HepG2 cells revealed multiple DNA loops within this region and two G4 structures marked with enhancer activity (H3K27Ac) residing in the center of the DNA interaction network (Figure 3.5A). In contrast, G4 structures are depleted in these regions in K562 cells, which are accompanied with much less RNAPII-mediated DNA interaction network in these

regions in K562 cells than HepG2 cells (Figure 3.5A**)**. Such analysis suggests that G4s may play a critical role in 3D genome architecture and modulate the expression of *AKR1C1-3* genes through long-range DNA interactions.

After PDS treatment, RNAPII exhibited markedly diminished occupancy at the two G4 regions, which is accompanied with reduced RNAPII-mediated DNA interactions as shown in the HiChIP-seq results (Figure 3.5A**)**. To further validate these findings, we conducted 3C-qPCR assay, which can accurately determine interaction frequencies between genomic loci. One interaction of interest is between the promoter of *AKR1C1* and its upstream enhancer (E-P interaction, Figure 3.5B). We measured ligation efficiencies between the constant fragment (with primer 1f) located in the promoter and five candidate fragments (with primer 2f-6f). Our results showed a markedly elevated ligation efficiency between 1f and 4f, which represents the physical interaction of *AKR1C1* promoter and upstream G4-containing enhancer in HepG2 cells under mock-treatment conditions. In accordance with HiChIP-seq results, we observed a significant decrease in ligation efficiency after PDS treatment, supporting an active role of G4 structure in this E-P interaction. Previous studies demonstrated that JQ1, a small-molecule, can specifically dislodge BRD4 from enhancers thereby dissolving mediator and RNAPII clusters (50,51). Treatment of JQ1 can cause reconfiguration of chromatin structure in selected gene loci (52). We observed a significantly diminished ligation efficiency in the same region following JQ1 treatment, indicating that the interaction observed between the transcription start site (TSS) region of *AKR1C1* and the distal G4 is promoter-enhancer interaction.

We also validated another interaction between two G4 structures in *AKR1C1-3* region. These two G4 structures serve as hubs in connecting multiple genomic loci and overlap with H3K27Ac enhancer marks. Because of the long distance between these two regions (~120 kb), a relative lower ligation frequency was observed compared to the aforementioned E-P interaction

(Figure 3.5C). However, 3C-qPCR assay showed that PDS treatment resulted in a significant attenuation in interaction between the two G4 loci (4f and 4g) compared to mock treatment (Figure 3.5C). Such diminution was also observed in cells treated with JQ1 (Figure 3.5C). These results demonstrated a physical linkage between two G4s involved in an enhancer-enhancer (E-E) interaction.

Next we examined the role of G4-dependent RNAPII-mediated DNA loops in the expression of *AKR1C1-3* genes. In accordance with the diminished RNAPII-mediated DNA interactions in this region, we observed attenuated expressions of *AKR1C1-3* genes in HepG2 cells (Figure 3.5D). By contrast, we did not observe any significant changes in expression of *AKR1C1-3* genes in K562 cells after PDS treatment, which is consistent with the lack of G4 structure and DNA loops in these regions in K562 cells (Figure 3.5E). The above data support a role of G4 structures in augmenting the occupancy of RNAPII in enhancer regions to stimulate transcription of target genes brought to close proximity by DNA looping. These results further substantiate the roles of G4 structures in cell type-specific RNAPII-mediated DNA looping and transcription regulation.

3.4. Conclusions

Under physiological conditions, G-rich regions of DNA can fold into G4 structures, which regulate important cellular processes including transcription. DNA G4 was first found to be involved in gene regulation by Hurley and coworkers (53), who observed that treatment of Burkitt's lymphoma cells with G4 ligands, e.g., PDS and TMPyP4, led to diminished transcription of *MYC* gene, whose promoter contains G4-forming sequence. With the availability of a G4 structure-specific antibody (i.e., BG4), recent studies unveiled an association between G4 and transcription regulation (12,14,54-56). For instance, overlapping analysis of DNMT1 ChIP-seq revealed its

significant enrichment at cellular G4 sites, which also exhibit much lower CpG methylation (19). A sequestration model, where the recruitment of DNMT1 to G4 inhibits its enzymatic activity and results in hypomethylated regions, was proposed to account for the role of G4 in modulating gene expression. A later study showed that different G4 folding states, measured by G4 ChIP-seq, is associated with distinct transcriptome profiles in two cell lines (54). Therefore, it is of interest to investigate the detailed mechanism through which DNA G4 structures modulate transcription.

Aside from promoters, transcription can also be modulated by distal regulatory elements like enhancers, which are remote from transcription start sites of target genes in the primary sequence, but close in 3D genome organization. Mediated by transcription factors and cofactors, E-P interactions initiate and promote RNAPII-mediated transcription (21). Integrative analysis showed significant enrichment of G4 at the TAD boundaries, which are proposed to be the structural scaffolds for E-P contacts. A recent study by Li *et al.* (13) revealed the ability of YY1, a transcription factor known to enable DNA looping (27), in binding with G4 structures *in vitro* and in cells, and found that disruption of YY1-G4 binding led to a diminution in YY1-mediated DNA looping.

With the encouraging results of YY1 transcription factor, we sought to investigate further how G4 functions in RNAPII-linked long-range DNA interactions and affects transcription in general. First, we employed bioinformatic analysis by comparing POLR2A ChIA-PET and BG4 ChIP-seq in three cell lines (i.e., HepG2, HEK293T and K562). We found a strong overlap, > 60% in HepG2 cells, between G4 structure sites and RNAPII-mediated DNA loops (Figure 3.1A). Our finding is consistent with previous integrative analysis of G4 ChIP-seq with Hi-C dataset (26). Moreover, we analyzed the association between interaction frequency and the presence of G4 structure, and found that more interactions are observed at cellular G4 loci (Figure 3.1D). Shuffling calculation at DNase I hypersensitive sites and sites with OQS substantiated our finding that G4

structures are important determinants for long-range DNA contacts. As DNA loops vary in different cell lines, we demonstrated that distinct DNA looping patterns are strongly associated with the cell type-specific distributions of G4 structures in chromatin. Together, our bioinformatic analysis lent evidence to support that G4 is involved in high-order chromatin organization and in RNAPII-mediated long-range DNA interactions. It is worth noting that there is so far no evidence supporting that RNAPII can bind directly with DNA G4 structures. However, many transcription factors exhibit ability in binding directly with G4 DNA (13,16,57-59). As noted above, one of these transcription factors, YY1, could bind to G4 DNA at low nM binding affinity, and this binding contributes to YY1-mediated DNA looping (13). It will be important to examine how other G4-binding transcription factors contribute to RNAPII-mediated DNA looping.

We also found that PDS, a small-molecule G4-binding ligand, could disrupt global RNAPII binding with a ~68% decrease in significant binding sites (Figure 3.2A). Specifically, those RNAPII-binding loci with G4 structures displayed more pronounced decreases following PDS treatment compared to those without (Figure 3.2B, C). As RNAPII constitutes the core component of the mammalian transcription machinery, our data suggest an important role of G4 in RNAPII-mediated long-range DNA interaction and transcription regulation. Importantly, by using HiChIP-seq, we detected attenuated RNAPII-mediated long-range DNA interactions following PDS treatment (Figure 3.3A, B**),** and we demonstrated that such diminishing effect was more pronounced in DNA loops with G4-containing anchors than those without (Figure 3.3C).

G4 structure has been proposed to play important roles in transcription regulation (5,11,19); however, limited studies demonstrated experimentally how distal G4s modulate gene expression (13). We combined RNA-seq with long-range DNA interaction data (ChIA-PET/HiChIP-seq) to unravel the regulatory roles of G4-depedent DNA loops in transcription. Transcriptome abundance profiling revealed higher expression of not only those genes with G4 in promoter regions but also

those connected to a distal G4 through RNAPII-mediated long-range DNA interactions (Figure 3.4B). In addition, PDS-induced alterations in expression of those genes with G4-dependent DNA loops are more pronounced than those genes connected with DNA loops lacking G4 structures (Figure 3.4C).

We further evaluated how G4-dependent DNA loops in a specific genomic region and how they modulate the expression of their target genes. Aldo-keto reductases family 1C (AKR1C) are a group of enzymes responsible for steroid reductions (60). AKR1C3 was shown to have an important role in the progression of prostate cancer (61) and several selective inhibitors of AKR1C3 have shown anti-tumor activity (62-64). Furthermore, a bioinformatic analysis showed elevated *AKR1C1-3* expression in liver cancer samples compared with normal liver samples (65). Poorer survival rate was observed in those cancer patients with high expression of *AKR1C1-3*, suggesting that they may serve as prognostic markers for liver cancer (65). Our HiChIP experiments demonstrated that treatment with a G4-binding ligand led to diminished RNAPII-mediated DNA loops (Figure 3.5A). By using 3C-qPCR assay, we validated our findings made from RNAPII HiChIP-Seq and demonstrated the participation of G4 structure in both the promoter-enhancer and enhancer-enhancer interactions in *AKR1C1-3* regions (Figure 3.5B, C). We further demonstrated that G4-dependent RNAPII-mediated DNA loops play an important role in regulating the expression of *AKR1C1-3* in HepG2 cells (Figure 3.5D, E). Considering the possible role of *AKR1C* family in liver cancer, our study in HepG2 cells provided an important knowledge basis for potential therapeutic interventions of liver cancer. In addition, we failed to observe any apparent impact of PDS treatment on the expression of the *AKR1C1-3* genes in K562 cells, which is in line with the lack of G4 structure-mediated DNA looping at these genetic loci in K562 cells. Our work, hence, also underscores a role of the interplay of G4 structure and RNAPII-mediated DNA looping in cell type-dependent gene expression in human cells.

In summary, we revealed, using a combination of bioinformatic and experimental approaches, that DNA G4 actively participates in RNAPII-mediated long-range DNA interactions (Figure 3.6). We also found that G4 structures not only locally modulate transcription in promoter regions, but also remotely regulate gene expression through long-range DNA interactions. Moreover, our work revealed a role of G4 structure in differentially modulating RNAPII-mediated DNA looping and expression of target genes in two different cell lines, which could stimulate future studies about the role of G4-dependent DNA loops in cell type-specific gene expression and in cancer biology. Together, our study provided new insights into the functional interplay of G4 structures and 3D genome architecture in regulating gene expression.

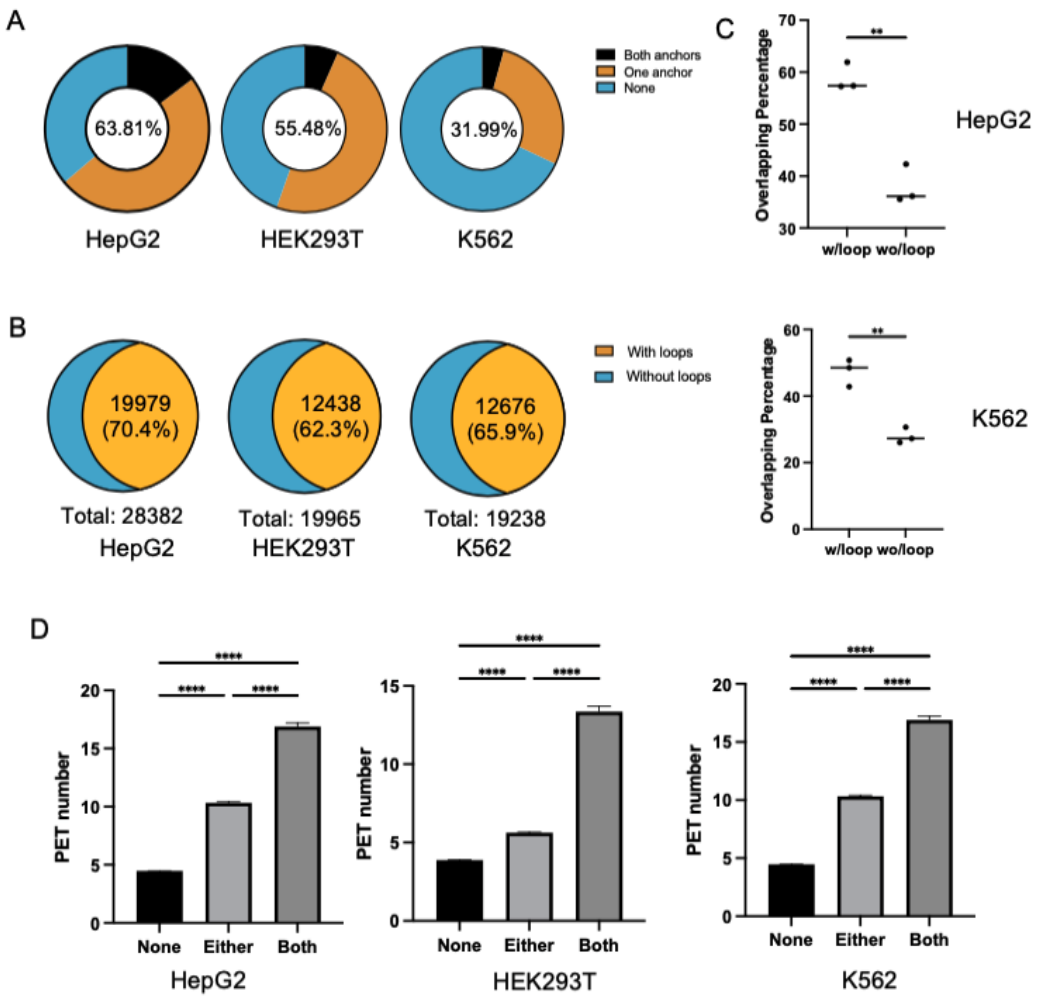

Figure 3.1. Overlapping analysis between RNAPII ChIA-PET with BG4 ChIP-seq peaks in HepG2, HEK293T and K562 cells. (A) Percentages of DNA loop anchors, as revealed from RNAPII ChIA-PET analysis, that overlap with G4 structure loci, as determined from BG4 ChIP-seq analysis. The ChIA-PET loop anchors are divided into three groups, with both anchors having G4 structures (Both anchors), only one of them having G4 structure (One anchor), or neither having G4 structures (None). (B) Percentages of G4 structure sites (obtained from BG4 ChIP-Seq) that overlap with anchors of RNAPII-mediated DNA loops (obtained from ChIA-PET analysis). (C) The percentages of G4 structure peaks overlapping with those RNAPII-binding sites that are involved with DNA loop formation vs. those that are not. Two-tailed Student's t-test with Welch's Correction, **, $p < 0.01$. (D) Statistical analysis of PET number of DNA loops with regard to anchor's overlapping with BG4 ChIP-seq peaks; shown are mean ± SEM. One-way ANOVA test, ****, $p < 0.0001$.

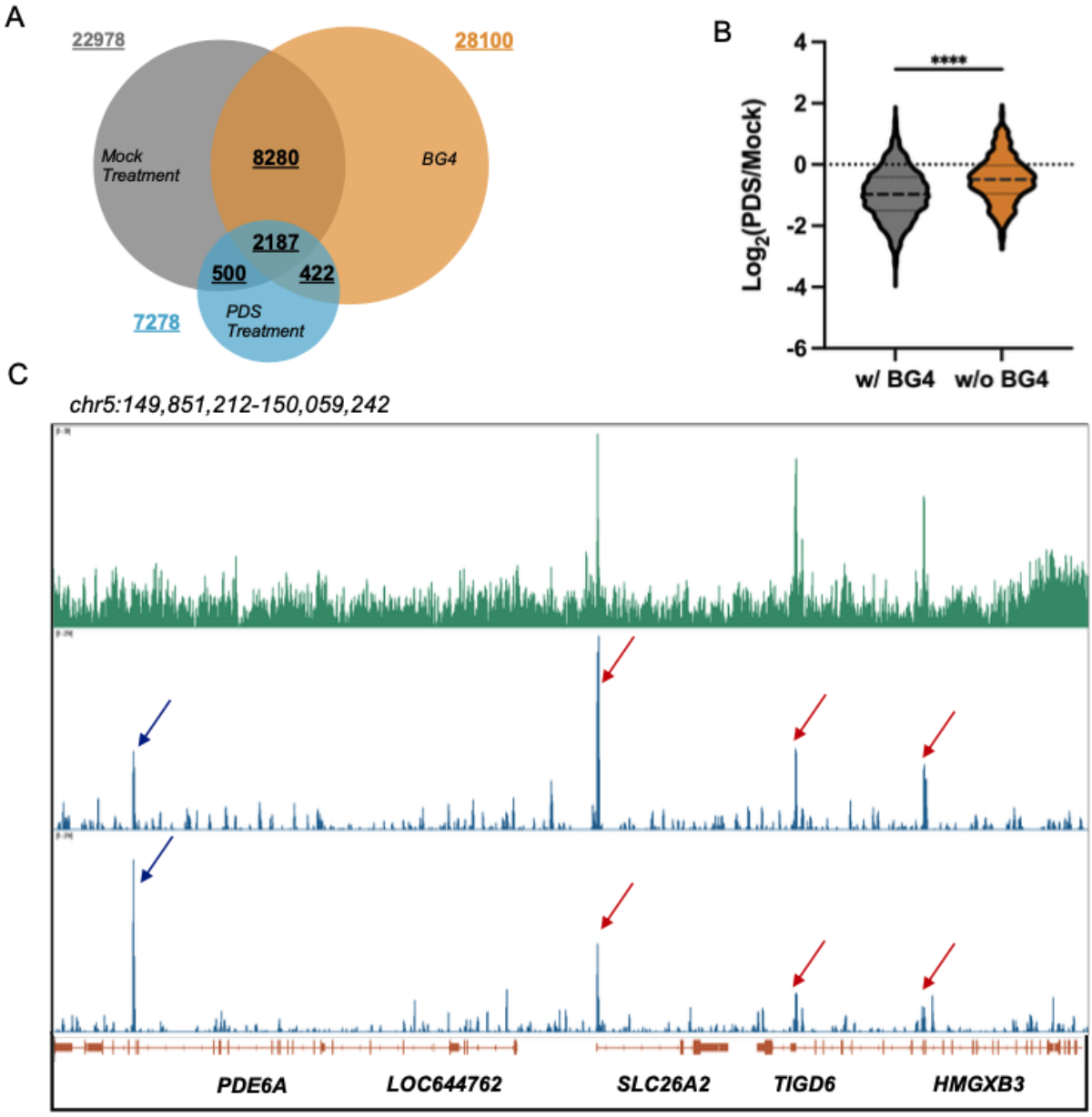

Figure 3.2. PDS treatment modulates genome-wide landscape of RNAPII occupancy. (A) A Venn diagram displaying the overlaps of RNAPII peaks in HepG2 cells that are mock- or PDS-treated, as revealed from HiChIP-Seq analysis, with BG4 ChIP-Seq peaks detected in HepG2 cells. (B) The ratios of RNAPII ChIP-Seq signal in PDS- over mock-treated HepG2 cells for those peaks that overlap (w/ BG4) or not (w/o BG4) with BG4 ChIP-Seq peaks. Two-tailed Student's t-test with Welch's Correction, ****, $p < 0.0001$. (C) IGV plots depicting diminished RNAPII ChIP signal at G4 structure loci following PDS treatment.

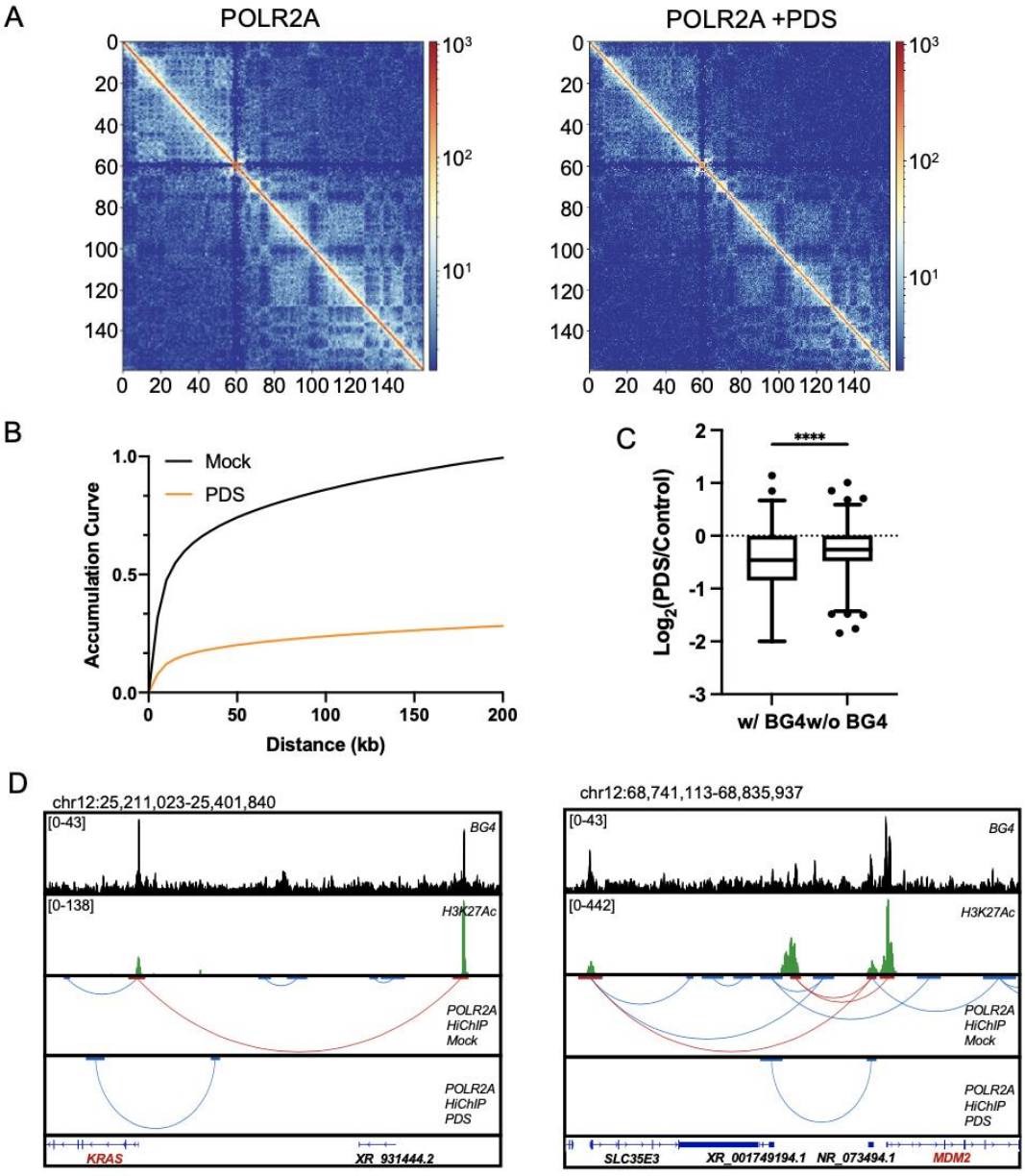

Figure 3.3. HiChIP-seq analysis showing that PDS preferentially disrupts RNAPII-linked long-range DNA interactions involving G4 structure loci. (A) HiChIP interaction matrices of RNAPII in chromosome 7 in HepG2 cells that were mock-treated (left) or treated with P PDS (right); (B) Aggregation analysis of RNAPII-mediated long-range DNA interactions in mock- and PDS-treated HepG2 cells; (C) HiChIP PET ratios in PDS- over mock-treated HepG2 cells with respect to overlap with BG4 ChIP-seq peaks. Two-tailed Student's t-test with Welch's Correction, ****, $p < 0.0001$. (D) POLR2A HiChIP-seq results for G4-mediated long-range DNA interactions involving the promoters of KRAS and MDM2 genes in mock- and PDS-treated HepG2 cells.

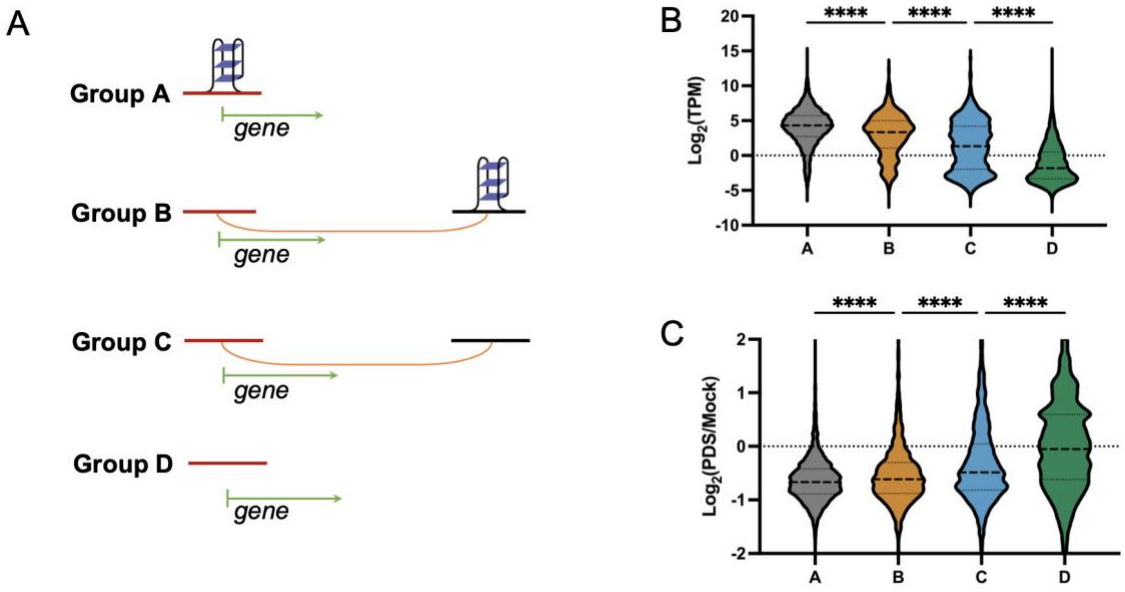

Figure 3.4. Consolidation analysis of RNA-seq and 3D genome architecture mapping. (A) A scheme depicting the grouping strategy. Genes were divided into four groups based on their association with G4 structures: Group A genes have G4 structures in their promoters; Group B and C genes do not contain G4 structures in their promoters, but these promoters are connected, via RNAPII-mediated NDA looping, to distal sites with and without G4 structures, respectively. The rest genes were classified into Group D. (B) Transcriptome profiles of each group of genes in mock-treated HepG2 cells. (C) Statistical analysis of PDS-induced alterations of the transcriptome in the four groups of genes in HepG2 cells. One-way ANOVA test, ****, $p <$ 0.0001.

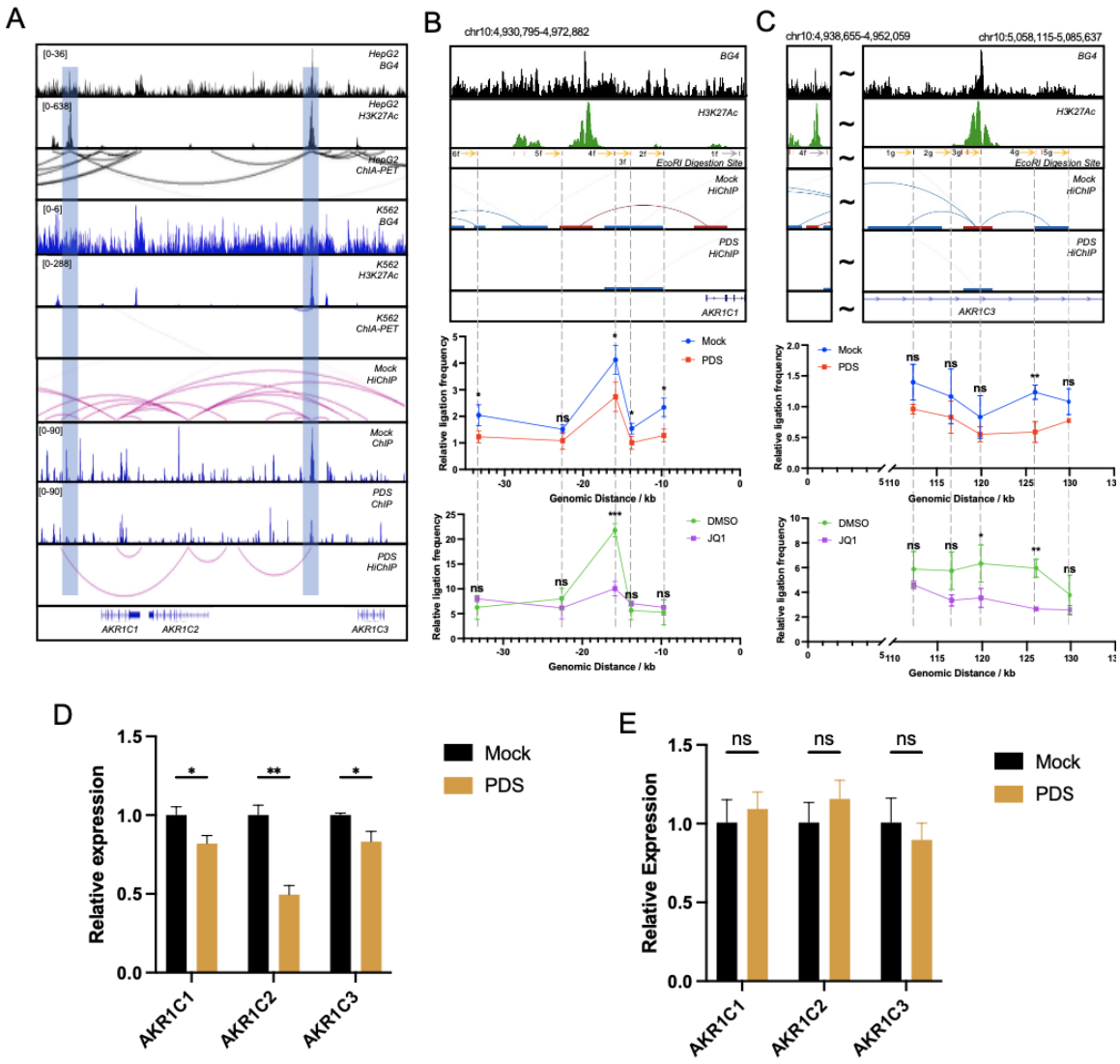

Figure 3.5. G4-dependent and RNAPII-linked DNA loops regulate the expression of AKR1C family genes. (A) RNAPII-mediated long-range DNA interactions within the regions of AKR1C1-3 gene in HepG2, but not in K562 cells. POLR2A HiChIP-seq results for the regions of AKR1C1-3 genes in mock- and PDS-treated HepG2 cells. (B-C) 3C-qPCR results for AKR1C1 E-P interaction and G4-mediated E-E interactions in HepG2 cells with or without PDS treatment, and with or without JQ1 treatment. The relative level of each ligation product was plotted according to its distance from the constant primer. The data were normalized to ERCC3 control interaction frequencies for each ligation product. The data represent mean± SD (n = 3). Two-tailed Student's t-test, *, $0.01 \leq p < 0.05$; **, $0.001 \leq p < 0.01$; ***, $0.0001 \leq p < 0.001$. (D-E) RT-qPCR (Mean ± SD, n = 3) results showing the relative expression levels of AKR1C1-3 genes in HepG2 and K562 cells with or without PDS treatment. Two-tailed Student's t-test with Benjamini and Hochberg correction for multiple comparison. *, $0.01 \leq p < 0.05$; **, $0.001 \leq p < 0.01$.

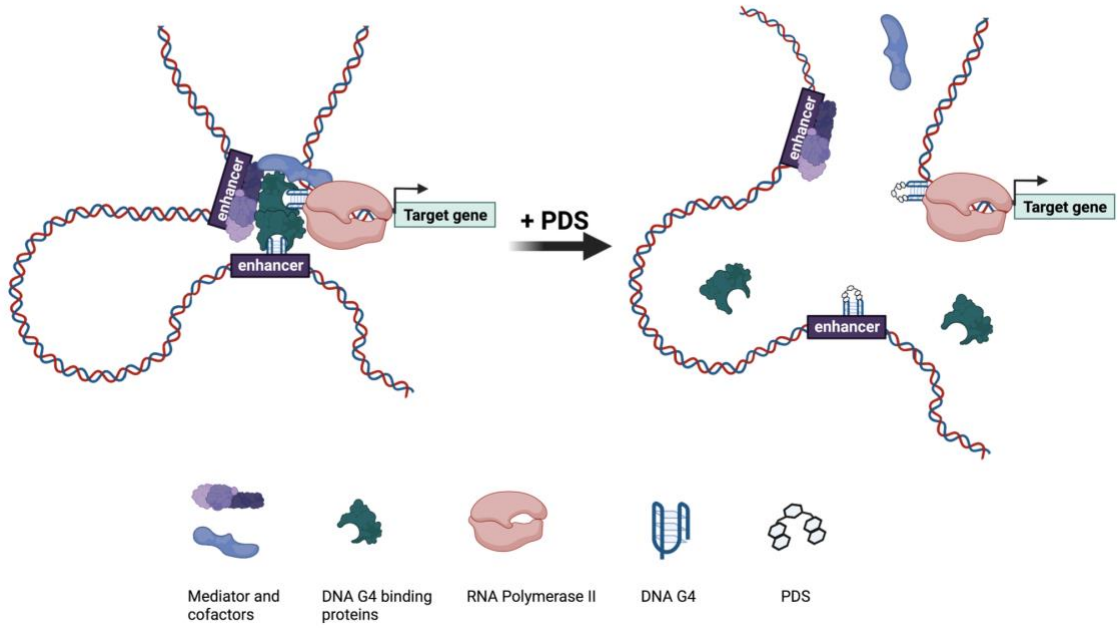

Figure 3.6. A model illustrating the involvement of G4 structures in RNAPII-linked long-range DNA interactions and in gene expression regulation. A small-molecule G4 ligand, PDS, can perturb G4-binding capacity of proteins (e.g., YY1) and disrupt 3D genome architecture.

3.5. References

1. Bochman, M.L., Paeschke, K. and Zakian, V.A. (2012) DNA secondary structures: stability and function of G-quadruplex structures. *Nat. Rev. Genet.*, **13**, 770-780.

2. Burge, S., Parkinson, G.N., Hazel, P., Todd, A.K. and Neidle, S. (2006) Quadruplex DNA: sequence, topology and structure. *Nucleic Acids Res.*, **34**, 5402-5415.

3. Hardin, C.C., Watson, T., Corregan, M. and Bailey, C. (1992) Cation-dependent transition between the quadruplex and Watson-Crick hairpin forms of d(CGCG3GCG). *Biochemistry*, **31**, 833-841.

4. Biffi, G., Tannahill, D., McCafferty, J. and Balasubramanian, S. (2013) Quantitative visualization of DNA G-quadruplex structures in human cells. *Nat. Chem.*, **5**, 182-186.

5. Hansel-Hertsch, R., Beraldi, D., Lensing, S.V., Marsico, G., Zyner, K., Parry, A., Di Antonio, M., Pike, J., Kimura, H., Narita, M. *et al.* (2016) G-quadruplex structures mark human regulatory chromatin. *Nat. Genet.*, **48**, 1267-1272.

6. Hansel-Hertsch, R., Spiegel, J., Marsico, G., Tannahill, D. and Balasubramanian, S. (2018) Genome-wide mapping of endogenous G-quadruplex DNA structures by chromatin immunoprecipitation and high-throughput sequencing. *Nat. Protoc.*, **13**, 551-564.

7. Parkinson, G.N., Lee, M.P. and Neidle, S. (2002) Crystal structure of parallel quadruplexes from human telomeric DNA. *Nature*, **417**, 876-880.

8. Fouquerel, E., Parikh, D. and Opresko, P. (2016) DNA damage processing at telomeres: The ends justify the means. *DNA Repair (Amst.)*, **44**, 159-168.

9. Valton, A.L. and Prioleau, M.N. (2016) G-Quadruplexes in DNA Replication: A Problem or a Necessity? *Trends Genet.*, **32**, 697-706.

10. Cogoi, S., Ferino, A., Miglietta, G., Pedersen, E.B. and Xodo, L.E. (2018) The regulatory G4 motif of the Kirsten ras (KRAS) gene is sensitive to guanine oxidation: implications on transcription. *Nucleic Acids Res.*, **46**, 661-676.

11. Robinson, J., Raguseo, F., Nuccio, S.P., Liano, D. and Di Antonio, M. (2021) DNA G-quadruplex structures: more than simple roadblocks to transcription? *Nucleic Acids Res.*, **49**, 8419-8431.

12. Fleming, A.M. and Burrows, C.J. (2020) Interplay of Guanine Oxidation and G-Quadruplex Folding in Gene Promoters. *J. Am. Chem. Soc.*, **142**, 1115-1136.

13. Li, L., Williams, P., Ren, W., Wang, M.Y., Gao, Z., Miao, W., Huang, M., Song, J. and Wang, Y. (2021) YY1 interacts with guanine quadruplexes to regulate DNA looping and gene expression. *Nat. Chem. Biol.*, **17**, 161-168.

14. Spiegel, J., Cuesta, S.M., Adhikari, S., Hansel-Hertsch, R., Tannahill, D. and Balasubramanian, S. (2021) G-quadruplexes are transcription factor binding hubs in human chromatin. *Genome Biol.*, **22**, 117.

15. Cogoi, S., Paramasivam, M., Membrino, A., Yokoyama, K.K. and Xodo, L.E. (2010) The KRAS promoter responds to Myc-associated zinc finger and poly(ADP-ribose) polymerase 1 proteins, which recognize a critical quadruplex-forming GA-element. *J. Biol. Chem.*, **285**, 22003-22016.

16. Raiber, E.A., Kranaster, R., Lam, E., Nikan, M. and Balasubramanian, S. (2012) A non-canonical DNA structure is a binding motif for the transcription factor SP1 in vitro. *Nucleic Acids Res.*, **40**, 1499-1508.

17. Pavlova, II, Tsvetkov, V.B., Isaakova, E.A., Severov, V.V., Khomyakova, E.A., Lacis, I.A., Lazarev, V.N., Lagarkova, M.A., Pozmogova, G.E. and Varizhuk, A.M. (2020) Transcription-facilitating histone chaperons interact with genomic and synthetic G4 structures. *Int. J. Biol. Macromol.*, **160**, 1144-1157.

18. Williams, P., Li, L., Dong, X. and Wang, Y. (2017) Identification of SLIRP as a G Quadruplex-Binding Protein. *J. Am. Chem. Soc.*, **139**, 12426-12429.

19. Mao, S.Q., Ghanbarian, A.T., Spiegel, J., Martinez Cuesta, S., Beraldi, D., Di Antonio, M., Marsico, G., Hansel-Hertsch, R., Tannahill, D. and Balasubramanian, S. (2018) DNA G-quadruplex structures mold the DNA methylome. *Nat. Struct. Mol. Biol.*, **25**, 951-957.

20. Bonev, B. and Cavalli, G. (2016) Organization and function of the 3D genome. *Nat. Rev. Genet.*, **17**, 661-678.

21. Schoenfelder, S. and Fraser, P. (2019) Long-range enhancer-promoter contacts in gene expression control. *Nat. Rev. Genet.*, **20**, 437-455.

22. Marchal, C., Sima, J. and Gilbert, D.M. (2019) Control of DNA replication timing in the 3D genome. *Nat. Rev. Mol. Cell Biol.*, **20**, 721-737.

23. Zheng, H. and Xie, W. (2019) The role of 3D genome organization in development and cell differentiation. *Nat. Rev. Mol. Cell Biol.*, **20**, 535-550.

24. Krijger, P.H. and de Laat, W. (2016) Regulation of disease-associated gene expression in the 3D genome. *Nat. Rev. Mol. Cell Biol.*, **17**, 771-782.

25. Kempfer, R. and Pombo, A. (2020) Methods for mapping 3D chromosome architecture. *Nat. Rev. Genet.*, **21**, 207-226.

26. Hou, Y., Li, F., Zhang, R., Li, S., Liu, H., Qin, Z.S. and Sun, X. (2019) Integrative characterization of G-Quadruplexes in the three-dimensional chromatin structure. *Epigenetics*, **14**, 894-911.

27. Weintraub, A.S., Li, C.H., Zamudio, A.V., Sigova, A.A., Hannett, N.M., Day, D.S., Abraham, B.J., Cohen, M.A., Nabet, B., Buckley, D.L. *et al.* (2017) YY1 Is a Structural Regulator of Enhancer-Promoter Loops. *Cell*, **171**, 1573-1588 e1528.

28. Davis, C.A., Hitz, B.C., Sloan, C.A., Chan, E.T., Davidson, J.M., Gabdank, I., Hilton, J.A., Jain, K., Baymuradov, U.K., Narayanan, A.K. *et al.* (2018) The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res.*, **46**, D794-D801.

29. Quinlan, A.R. and Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841-842.

30. Mumbach, M.R., Rubin, A.J., Flynn, R.A., Dai, C., Khavari, P.A., Greenleaf, W.J. and Chang, H.Y. (2016) HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nat. Methods*, **13**, 919-922.

31. Servant, N., Varoquaux, N., Lajoie, B.R., Viara, E., Chen, C.J., Vert, J.P., Heard, E., Dekker, J. and Barillot, E. (2015) HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.*, **16**, 259.

32. Ramirez, F., Bhardwaj, V., Arrigoni, L., Lam, K.C., Gruning, B.A., Villaveces, J., Habermann, B., Akhtar, A. and Manke, T. (2018) High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nat. Commun.*, **9**, 189.

33. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. and Gingeras, T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15-21.

34. Liao, Y., Smyth, G.K. and Shi, W. (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, **30**, 923-930.

35. Love, M.I., Huber, W. and Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.

36. Hagege, H., Klous, P., Braem, C., Splinter, E., Dekker, J., Cathala, G., de Laat, W. and Forne, T. (2007) Quantitative analysis of chromosome conformation capture assays (3C-qPCR). *Nat. Protoc.*, **2**, 1722-1733.

37. Sun, H., Wu, J., Wickramasinghe, P., Pal, S., Gupta, R., Bhattacharyya, A., Agosto-Perez, F.J., Showe, L.C., Huang, T.H. and Davuluri, R.V. (2011) Genome-wide mapping of RNA Pol-II promoter usage in mouse tissues by ChIP-seq. *Nucleic Acids Res.*, **39**, 190-201.

38. Consortium, E.P. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57-74.

39. Li, C., Wang, H., Yin, Z., Fang, P., Xiao, R., Xiang, Y., Wang, W., Li, Q., Huang, B., Huang, J. *et al.* (2021) Ligand-induced native G-quadruplex stabilization impairs transcription initiation. *Genome Res.*, **31**, 1546-1560.

40. Khuu, P., Sandor, M., DeYoung, J. and Ho, P.S. (2007) Phylogenomic analysis of the emergence of GC-rich transcription elements. *Proc. Natl. Acad. Sci. U. S. A.*, **104**, 16528-16533.

41. Chambers, V.S., Marsico, G., Boutell, J.M., Di Antonio, M., Smith, G.P. and Balasubramanian, S. (2015) High-throughput sequencing of DNA G-quadruplex structures in the human genome. *Nat. Biotechnol.*, **33**, 877-881.

42. Sanyal, A., Lajoie, B.R., Jain, G. and Dekker, J. (2012) The long-range interaction landscape of gene promoters. *Nature*, **489**, 109-113.

43. Ross-Innes, C.S., Stark, R., Teschendorff, A.E., Holmes, K.A., Ali, H.R., Dunning, M.J., Brown, G.D., Gojis, O., Ellis, I.O., Green, A.R. *et al.* (2012) Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature*, **481**, 389-393.

44. Lareau, C.A. and Aryee, M.J. (2018) diffloop: a computational framework for identifying and analyzing differential DNA loops from sequencing data. *Bioinformatics*, **34**, 672-674.

45. Rodriguez, R., Muller, S., Yeoman, J.A., Trentesaux, C., Riou, J.F. and 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.

46. He, X., Yuan, J. and Wang, Y. (2021) G3BP1 binds to guanine quadruplexes in mRNAs to modulate their stabilities. *Nucleic Acids Res.*, **49**, 11323-11336.

47. Lago, S., Nadai, M., Ruggiero, E., Tassinari, M., Marusic, M., Tosoni, B., Frasson, I., Cernilogar, F.M., Pirota, V., Doria, F. *et al.* (2021) The MDM2 inducible promoter folds into four-tetrad antiparallel G-quadruplexes targetable to fight malignant liposarcoma. *Nucleic Acids Res.*, **49**, 847-863.

48. Balasubramanian, S., Hurley, L.H. and Neidle, S. (2011) Targeting G-quadruplexes in gene promoters: a novel anticancer strategy? *Nat. Rev. Drug Discov.*, **10**, 261-275.

49. Buratowski, S. (1994) The basics of basal transcription by RNA polymerase II. *Cell*, **77**, 1-3.

50. Cho, W.K., Spille, J.H., Hecht, M., Lee, C., Li, C., Grube, V. and Cisse, II. (2018) Mediator and RNA polymerase II clusters associate in transcription-dependent condensates. *Science*, **361**, 412-415.

51. Kanno, T., Kanno, Y., LeRoy, G., Campos, E., Sun, H.W., Brooks, S.R., Vahedi, G., Heightman, T.D., Garcia, B.A., Reinberg, D. *et al.* (2014) BRD4 assists elongation of both coding and enhancer RNAs by interacting with acetylated histones. *Nat. Struct. Mol. Biol.*, **21**, 1047-1057.

52. Choi, H.I., An, G.Y., Yoo, E., Baek, M., Binas, B., Chai, J.C., Lee, Y.S., Jung, K.H. and Chai, Y.G. (2022) The bromodomain inhibitor JQ1 up-regulates the long non-coding RNA MALAT1 in cultured human hepatic carcinoma cells. *Sci. Rep.*, **12**, 7779.

53. Siddiqui-Jain, A., Grand, C.L., Bearss, D.J. and 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.

54. Lago, S., Nadai, M., Cernilogar, F.M., Kazerani, M., Dominiguez Moreno, H., Schotta, G. and Richter, S.N. (2021) Promoter G-quadruplexes and transcription factors cooperate to shape the cell type-specific transcriptome. *Nat. Commun.*, **12**, 3885.

55. Zheng, K.W., Zhang, J.Y., He, Y.D., Gong, J.Y., Wen, C.J., Chen, J.N., Hao, Y.H., Zhao, Y. and Tan, Z. (2020) Detection of genomic G-quadruplexes in living cells using a small artificial protein. *Nucleic Acids Res.*, **48**, 11706-11720.

56. Di Antonio, M., Ponjavic, A., Radzevicius, A., Ranasinghe, R.T., Catalano, M., Zhang, X., Shen, J., Needham, L.M., Lee, S.F., Klenerman, D. *et al.* (2020) Single-molecule visualization of DNA G-quadruplex formation in live cells. *Nat. Chem.*, **12**, 832-837.

57. Simko, E.A.J., Liu, H., Zhang, T., Velasquez, A., Teli, S., Haeusler, A.R. and Wang, J. (2020) G-quadruplexes offer a conserved structural motif for NONO recruitment to NEAT1 architectural lncRNA. *Nucleic Acids Res.*, **48**, 7421-7438.

58. Lyonnais, S., Tarres-Sole, A., Rubio-Cosials, A., Cuppari, A., Brito, R., Jaumot, J., Gargallo, R., Vilaseca, M., Silva, C., Granzhan, A. *et al.* (2017) The human mitochondrial transcription factor A is a versatile G-quadruplex binding protein. *Sci. Rep.*, **7**, 43992.

59. Li, L., Williams, P., Gao, Z. and Wang, Y. (2020) VEZF1-guanine quadruplex DNA interaction regulates alternative polyadenylation and detyrosinase activity of VASH1. *Nucleic Acids Res.*, **48**, 11994-12003.

60. Zeng, C.M., Chang, L.L., Ying, M.D., Cao, J., He, Q.J., Zhu, H. and Yang, B. (2017) Aldo-Keto Reductase AKR1C1-AKR1C4: Functions, Regulation, and Intervention for Anti-cancer Therapy. *Front. Pharmacol.*, **8**, 119.

61. Adeniji, A.O., Chen, M. and Penning, T.M. (2013) AKR1C3 as a target in castrate resistant prostate cancer. *J. Steroid Biochem. Mol. Biol.*, **137**, 136-149.

62. Endo, S., Oguri, H., Segawa, J., Kawai, M., Hu, D., Xia, S., Okada, T., Irie, K., Fujii, S., Gouda, H. *et al.* (2020) Development of Novel AKR1C3 Inhibitors as New Potential Treatment for Castration-Resistant Prostate Cancer. *J. Med. Chem.*, **63**, 10396-10411.

63. Liu, Y., He, S., Chen, Y., Liu, Y., Feng, F., Liu, W., Guo, Q., Zhao, L. and Sun, H. (2020) Overview of AKR1C3: Inhibitor Achievements and Disease Insights. *J. Med. Chem.*, **63**, 11305-11329.

64. Byrns, M.C., Jin, Y. and Penning, T.M. (2011) Inhibitors of type 5 17beta-hydroxysteroid dehydrogenase (AKR1C3): overview and structural insights. *J. Steroid Biochem. Mol. Biol.*, **125**, 95-104.

65. Zhao, S.F., Wang, S.G., Zhao, Z.Y. and Li, W.L. (2019) AKR1C1-3, notably AKR1C3, are distinct biomarkers for liver cancer diagnosis and prognosis: Database mining in malignancies. *Oncol. Lett.*, **18**, 4515-4522.

Chapter 4: G3BP1 Binds to Guanine Quadruplexes in mRNAs to Modulate Their Stabilities

4.1. Introduction

The RNA guanine quadruplexes (rG4s) are non-canonical four-stranded RNA structures that form in guanine (G)-rich regions of the transcriptome (1). rG4s comprise two or more stacked layers of G-quartets, in which four guanines are assembled in a planar configuration via Hoogsteen hydrogen bonding and are stabilized by K^+ ions (1). *In vitro* rG4 sequencing (rG4-seq) and *in silico* rG4 prediction revealed over 13,000 potential rG4-forming sites in the human transcriptome (2-5). Although these rG4s were suggested to be largely unfolded in cells (3), immunofluorescence microscopy analysis using G4 structure-specific antibody, live-cell imaging with rG4-binding fluorescent probes, and live-cell RNA labeling with N_3-kethoxal followed by deep sequencing all support the existence of rG4s in cells (6-9). Hence, rG4 structures might be highly dynamic in cells.

Bioinformatic predictions and rG4-seq results revealed that rG4s are enriched within 5′- and 3′-untranslated regions (UTRs) of mRNAs, which may constitute an important mechanism for post-transcriptional regulations of gene expression (2,3,10). Indeed, rG4s are implicated in the control of mRNA targeting, processing, translation and degradation (11-14). The biological functions of rG4s in cells often involve cellular proteins (1,15,16). For instance, DHX36 unwinds rG4 structures located in 3′-UTRs of mRNAs to prevent their accumulation in stress granules (17). Thus, understanding fully the biological functions of rG4s entails the identification and functional assessments of rG4-recognition proteins.

In recent years, various methods have been developed for the identification of rG4-binding proteins. Most methods rely on affinity pull-down followed by mass spectrometric analysis (18-22). Several rG4-binding proteins have been identified so far, including AVEN, CNBP, DDX21, DDX3X, DHX36, FMRP, nucleolin, etc. (18,22-27). There are, however, likely many yet identified rG4-binding proteins.

The Encyclopedia of DNA Elements (ENCODE) project has produced hundreds of enhanced UV crosslinking and immunoprecipitation-sequencing (eCLIP-seq) datasets for RNA-binding proteins (RBPs) (28). We posit that the binding sites for rG4-binding proteins in the transcriptome should overlap extensively with rG4 sites identified from rG4-Seq. Herein, we employed a bioinformatic analysis to identify candidate rG4-binding proteins on the basis of similarity between RBP-binding sites and rG4 structure loci in the transcriptome, which are obtained from publicly available eCLIP-seq (29,30) and rG4-seq (2) datasets, respectively. We demonstrated that, among the many putative rG4-binding proteins identified, G3BP1 (Ras GTPase-activating protein-binding protein 1), a stress granule protein, could bind directly and selectively with rG4. We also revealed that this binding modulates the stabilities of mRNAs bearing rG4 structures in the UTRs.

4.2 Materials and Methods

Cell Culture

HeLa and HEK293T (293T) cells, which were purchased from ATCC (Manassas, VA), were cultured in Dulbecco's modified Eagle's medium (DMEM, life Technologies) containing 10% fetal bovine serum (Invitrogen) and 1% penicillin and streptomycin (Invitrogen), and the cells were maintained at 37°C in an incubator containing 5% CO_2.

Bioinformatic Analysis

ENCODE data were retrieved from the ENCODE portal under eCLIP assay title and cell line biosample classification. A total of 223 experimental results were downloaded and the merged narrowpeak files were employed for overlapping analysis. rG4-seq data of HeLa cells were obtained using GEO accession number GSE77282 (2). IntervalStats (31) was employed for overlapping analysis. hg19 genome, rG4-KPDS-hit peak, and merged narrowpeak files of RBP eCLIP-seq peaks were used as domain, reference, and query, respectively. The resulting output was further filtered with a *p*-value cutoff of 0.05. Overlapping percentage was calculated as (# of overlapped peaks)/(total # of peaks for the target protein)×100%. Binding motif and peak annotation were analyzed by using HOMER (v4.11) (32). Signal enrichment was analyzed by using bwtool (33). The mapped reads were visualized using the Integrative Genomics Viewer (IGV_2.6.0) (34).

Purification of Recombinant Proteins

The plasmid for producing recombinant GST-G3BP1 was constructed by first amplifying the *G3BP1* gene from a cDNA library with primers containing BamHI and XhoI restriction recognition sites. The PCR amplicons were restriction digested and ligated into pGEX-4T1 vector, where the successful incorporation of the *G3BP1* coding sequence (CDS) was confirmed by sequencing. For truncated proteins, the corresponding CDS was amplified by PCR and inserted into the pGEX-4T1 vector using the same method.

The plasmids were transformed into competent Rosetta (DE3) pLysS *Escherichia coli* cells, and protein expression was induced by incubating cells with 1 mM isopropyl *β*-D-1-thio-galactopyranoside (IPTG, Sigma) at 16°C for 20 h. The cells were subsequently harvested by centrifugation and lysed by sonication in a 20-mL ice-cold PBS buffer containing 10% (v/v) glycerol and 1 mM phenylmethylsulfonyl fluoride (PMSF, Sigma) for 10 min. The cell lysate was then centrifuged at 10,000 *g* for 15 min. The GST-tagged proteins were purified from the

supernatant by using glutathione agarose beads (Pierce) following the manufacturer's recommended procedures. The full-length GST-G3BP1 protein was further purified using size-exclusion chromatography with a Superdex 200 increase 5/150 GL column and an AKTA Purifier 10 FPLC system (GE Healthcare). Protein purity was verified by SDS-PAGE analysis, quantified by Quick Start Bradford Protein Assay kit (Bio-Rad), and used immediately or stored at -80°C until use.

Fluorescence Anisotropy

Fluorescently labeled RNA probes (500 nM, Integrated DNA Technologies) were dissolved in an RNase-free buffer containing 10 mM Tris-HCl (pH 7.5), 100 mM KCl and 0.1 mM EDTA. The annealing was conducted by heating the solution to 95°C for 5 min, followed by cooling slowly to room temperature over 3 h. The binding assays were performed with 10 nM RNA probes and the indicated concentrations of recombinant G3BP1 protein in a 60-μL binding buffer containing 10 mM Tris-HCl (pH 7.5), 1 mM EDTA, 100 mM KCl, 0.1 mM DTT, and 10 μg/mL BSA. After incubation on ice for 30 min, fluorescence anisotropy was recorded on a Horiba QuantaMaster-400 spectrofluorometer (Photon Technology International), with the excitation and emission wavelengths being set at 550 and 580 nm, respectively, as described previously (35). The instrument G factor was determined before anisotropy measurements, and the K_d values were calculated with GraphPad Prism 8 software using non-linear regression for curve fitting with a one site-specific binding model.

Electrophoretic Mobility Shift Assay (EMSA)

EMSA was performed using a previously reported method with some modifications (36). Briefly, various concentrations of recombinant G3BP1 protein were incubated with 200 fmol of fluorescently labeled RNA probes in a binding buffer (10 mM Tris-HCl, pH 7.5, 1 mM EDTA, 100 mM KCl, 0.1 mM DTT, 10 μg/mL BSA). The mixtures were incubated on ice for 30 min and

the protein-bound RNA substrate was separated from free RNA on a 6% native polyacrylamide gel using 1× TAE (40 mM Tris-acetate, pH 8.0, 2 mM EDTA) by electrophoresis at 4°C. Electrophoresis was performed at 120 V for 22 min and the gel was imaged with an Odyssey Imaging System (LI-COR Biosciences).

Circular Dichroism (CD) Spectroscopy

The CD spectra for G3BP1 protein, and annealed PITX1 rG4 and NRAS rG4 (at 3 μM each) in an RNase-free buffer (10 mM Tris-HCl, pH 7.5, 100 mM KCl and 1 mM EDTA) were recorded separately in the wavelength range of 200-320 nm on a Jasco-815 spectropolarimeter. Additionally, a 3-μM solution of annealed PITX1 rG4 was incubated with an equal concentration of G3BP1 protein in the same RNase-free buffer at 4°C for 30 min, and the CD spectrum of mixture was subsequently acquired in the same wavelength range. The CD spectrum of G3BP1 protein was subtracted from the composite CD spectrum of the mixture to yield the CD spectrum of rG4 in the G3BP1-rG4 complex.

***In Vitro* Pull-down Experiment**

Biotin-conjugated PITX1 rG4 and rM4 were annealed individually in a buffer containing 10 mM Tris-HCl (pH 7.5), 100 mM KCl and 0.1 mM EDTA. The annealed RNA probes were incubated with high-capacity streptavidin agarose beads (Thermo Pierce) in a buffer containing 10 mM Tris-HCl (pH 7.5), 100 mM KCl, 1 mM EDTA and 0.1 mM DTT at 4°C for 1 h. After washing for three times with the binding buffer, the RNA-bound beads were incubated with the lysate of HEK293T cells at 4°C for 2 h in a buffer containing 10 mM Tris-HCl (pH 7.5), 100 mM KCl, 1 mM EDTA, 0.1 mM DTT, protease inhibitor (Sigma) and 0.4 units μL^{-1} RNase inhibitor (NEB). The beads were then washed thoroughly with the binding buffer supplemented with increasing concentrations of NaCl (100, 150 and 200 mM), followed by eluting the bound

proteins from the beads through boiling in 3 × SDS-PAGE loading buffer for 5 min. The resulting samples were subjected to Western blot analysis.

CRISPR/Cas9-mediated Ablation of *G3BP1* Gene

G3BP1$^{-/-}$ HEK293T and HeLa cells were generated by genome editing with the CRISPR/Cas9 system following the previously reported protocol (37), where the single guide RNA (sgRNA) was designed according to previously published method (38). The guide sequence for the production of sgRNA targeting *G3BP1* gene was inserted into the hSpCas9 plasmid pX330 (Addgene) at the BbsI digestion sites. After transfection and clonal isolation, successful deletion of the *G3BP1* gene in single-cell clones was screened by Western blot using anti-G3BP1 antibody and the deleted loci in genomic DNA was further identified by Sanger sequencing.

Western Blot

Protein samples were separated on a 10% SDS-PAGE gel and transferred onto a nitrocellulose membrane (Bio-Rad). After blocking with blotting-grade blocker (Bio-Rad), the membrane was incubated with PBS-T (PBS buffer with 0.05% Tween 20) containing primary antibody and 5% BSA for 2 h, and then incubated with the HRP-conjugated secondary antibody in a 5% blotting-grade blocker. Following thorough washing with PBS-T, the immunoblots were detected using ECL Western blotting detection reagent (Amersham). Primary antibodies used in this study included G3BP1 (MBL International, RN048PW; 1:2000), PITX1 (Proteintech, 10873-1-AP; 1:1000), KHSRP (Proteintech, 55409-1-AP; 1:500), ACTR2 (Proteintech, 10922-1-AP; 1:2000), FLAG-tag (Cell Signaling Technology, D6W5B; 1:2000), and GAPDH (Santa Cruz, sc-32233; 1:10000).

Real-time Quantitative PCR (RT-qPCR)

Total RNA was extracted using Omega Total RNA Kit I (Omega) and quantified. Reverse transcription was conducted using M-MLV Reverse Transcriptase (Promega) to obtain

the cDNA library. RT-qPCR was performed using Luna® Universal qPCR Master Mix (NEB) on the CFX96 RT-qPCR detection system (Bio-rad).

Dual-Luciferase Reporter Assay

The wild-type plasmid (PITX1-WT) for the reporter assay was constructed from PCR amplification of the *PITX1* 3′-UTR (1045 bp) with primers containing XbaI and FseI restriction recognition sites. The digested PCR product was ligated into pGL3-promoter vector (Promega) and the successful incorporation of the *PITX1* 3′-UTR was confirmed by sequencing. The corresponding plasmid with the quadruplex sequences being mutated (PITX1-3Qm) was constructed by site-directed mutagenesis (39), and the successful mutation was again verified by sequencing. The Flag-PITX1-WT and Flag-PITX1-3Qm plasmids were generated in two steps. First, the *PITX1* CDS (945 bp) was amplified by PCR with primers containing NotI and EcoRI restriction recognition sites, and the digested PCR product was ligated into a Flag-tagged pRK7 vector to get the pRK7-PITX1 plasmid. Subsequently, the Flag-PITX1 sequence was amplified from the pRK7-PITX1 plasmid with primers containing NcoI and XbaI restriction recognition sites, and the digested PCR product was ligated into the PITX1-WT and PITX1-3Qm reporter plasmids through replacing the coding sequence of the luciferase gene. The successful construction of Flag-PITX1-WT and Flag-PITX1-3Qm plasmids was confirmed by sequencing.

For the reporter assay, HEK293T cells and the isogenic *G3BP1*$^{-/-}$ cells were seeded in 12-well plates at a density of 2×10^5 cells per well. After 24 h, the cells (at ~50% confluency) were co-transfected with the 0.05 μg renilla luciferase plasmid (pRL-CMV, Promega) and 1 μg firefly luciferase plasmid (PITX1-WT or PITX1-3Qm). After another 12 h, the cells were treated with 20 μM pyridostatin (PDS) or mock-treated with water. The cells were harvested for measurements 12 h later. For luciferase reporter assay, the attached cells were lysed in 1× passive lysis buffer and vortexed to obtain a homogeneous cell lysate. The firefly and renilla luciferase activities of

the cell lysates were measured, with a 10-sec read time, using the dual-luciferase® reporter assay system and a luminometer (Promega), following the manufacturer's instructions. For mRNA expression detection, total RNA was extracted and again quantified using RT-qPCR.

RNA Half-life Measurements

HEK293T and the isogenic *G3BP1*$^{-/-}$ cells were seeded in 6-well plates at a density of 5 × 10^5 cells per well. After 24 h, the cells (at ~50% confluency) were untreated or treated with 20 μM PDS for 12 h before adding 5 μg/mL of actinomycin D to inhibit transcription. After the addition of actinomycin D, the cells were harvested at 0, 0.5, 1, 1.5, 3, 4.5, 6 and 9 h for total RNA extraction and the levels of transcripts of interest at different time points were quantified using RT-qPCR. The levels of target mRNAs at these time points were normalized to that of GAPDH and further normalized to that at 0 h to obtain the percentages of remaining mRNAs. RNA half-life was calculated with GraphPad Prism 8 software by fitting the percentages of remaining mRNA with single-phase exponential decay kinetics.

Single-end Enhanced Cross-linking and Immunoprecipitation-Sequencing (seCLIP-Seq) and Data Analysis

HEK293T cells were plated in 150 mm dishes at a density of 1 × 10^7 cells per dish. After 24 h, the cells (at ~60% confluency) were untreated or treated with 20 μM PDS for 12 h, followed by irradiating with UVC to induce protein-RNA cross-linking, and the subsequent seCLIP experiments (40) were performed with an eCLIP Library Prep Kit (Eclipse BioInnovations) following the manufacturer's instructions. Two biological replicates were conducted for each experiment, and 20 million cells were used for each replicate. seCLIP cDNA libraries were subsequently quantified using an Agilent 2100 Bioanalyzer and multiplexed for sequencing on an Illumina HiSeq 4000 Sequencing System with single-end 100-bp read length. Sequencing data were analyzed following the eCLIP-seq processing pipeline with the default setting (28). For the

comparison between “Ctrl” and “PDS” datasets, BEDtools was used with the criteria of at least 1 bp or 10 bp overlap for defining overlapped peaks (41). Metagene analysis was conducted using MetaPlotR Perl/R pipeline (42). G4 prediction was performed using G4Hunter, with the window size and threshold score being 25 and 1.4, respectively (43).

4.3 Results

Bioinformatic Discovery of Candidate rG4-binding Proteins

By employing a previously reported bioinformatic method for evaluating the similarity of chromatin immunoprecipitation-sequencing (ChIP-seq) data (44), we assessed, by calculating the *p*-values for proximity, the percentages of overlap between rG4 loci, which were uncovered by rG4-Seq (2), and the RNA-binding sites of 150 RBPs, which are based on the 233 eCLIP-seq datasets in the ENCODE database (29,30) (Figure 4.1**)**. To illustrate, we compared two RBPs with markedly different overlapping percentages with rG4-Seq data, i.e., AKAP1 (68.85%, a total of 5878 peaks) and MATR3 (10.32%, a total of 7168 peaks). Our results showed that most AKAP1 eCLIP-seq peaks overlap with rG4-seq peaks; MATR3 eCLIP-seq peaks, however, are staggered with rG4-seq peaks, suggesting the robustness of the bioinformatic approach in revealing overlaps between eCLIP-seq and rG4-seq peaks.

Importantly, the rG4-seq peaks exhibited high percentages of overlap with eCLIP-seq peaks of some known rG4-binding proteins, including DDX3X, FMR1, GRSF1, SRSF1 and YBX3 (Figure 4.1B) (18,20,45), demonstrating the feasibility of this method in discovering novel rG4-binding proteins. We also evaluated the overlap between our rG4-binding proteins and the published rG4-binding proteins, and found that ~50% of known rG4-binding proteins were identified in this study. Aside from these known rG4-binding proteins, our bioinformatic analysis led to the discovery of a large number of candidate new rG4-binding proteins (Figure 4.1B).

G3BP1 Is an rG4-binding Protein

Among these candidate rG4-binding proteins, G3BP1, a well-known stress granule protein, exhibits a 49.1% overlap in its eCLIP-seq peaks (3210 out of 6541) with rG4-seq peaks. Signal enrichment analysis also revealed strong overlap between G3BP1 and rG4-seq peaks (Figure 4.1C). Results from motif analysis of the overlapping peaks revealed that the most enriched motif contains a G-rich sequence satisfying the criteria for rG4 formation (46) (Figure 4.1D). In addition, two representative overlapping regions contain putative rG4-forming sequences, including the previously validated rG4 sequence in the 3′-UTR of LRP5 mRNA (47). Moreover, G3BP1 was recently shown to bind preferentially with mRNAs with highly structured 3′-UTRs (48). Together, these results suggest that G3BP1 is an rG4-binding protein, and we decided to choose this protein for further study.

We recognized that some proteins may bind to rG4 structures indirectly via protein-protein interactions, which may also give rise to high levels of overlap between their eCLIP-seq peaks and rG4-seq peaks. Hence, we next asked whether G3BP1 can bind directly with rG4 structures. To this end, we generated recombinant full-length G3BP1 protein and measured, by using fluorescence anisotropy, its binding affinities toward two previously characterized rG4 probes, one derived from the 5′-untranslated region (5′-UTR) of NRAS mRNA and the other from the 3′-UTR of PITX1 mRNA (13,49), and the corresponding mutated probes (rM4) that are unable to fold into G4 structure. CD measurement results showed that both PITX1 and NRAS rG4 sequences can assemble into parallel G4 topology, as manifested by negative and positive CD peaks at around 240 and 260 nm, respectively (50). The fluorescence anisotropy results showed that G3BP1 binds strongly with both rG4 probes, with the K_d values being 63 ± 10 and 60 ± 5 nM for rG4 structures derived from the mRNAs of *PITX1* and *NRAS* genes, respectively (Figure 4.2A-B). Moreover, the binding capacities of rG4 probes are much higher than those of the

corresponding rM4 probes, with the K_d values being 761 ± 73 and 266 ± 36 nM, respectively (Figure 4.2A-B).

We next assessed the binding capacities of G3BP1 toward rG4 and rM4 in cell lysate. To this end, we performed an *in vitro* pull-down experiment with biotin-labeled PITX1 rG4 and rM4 probes. The results showed that, with the use of the same amount of lysate, PITX1 rG4 probe was capable of pulling down >10-fold more G3BP1 than the corresponding rM4 probe under the same experimental conditions (Figure 4.2C). This is consistent with our fluorescence anisotropy results (Figure 4.2A), suggesting that G3BP1 is able to bind selectively to rG4 over rM4 in complex sample matrix (i.e., whole-cell protein lysate).

We also found that the G3BP1-rG4 interaction could be disrupted by PDS, a small-molecule ligand that can bind to and stabilize G4 structures (51). Additionally, CD measurement results showed that the G3BP1-rG4 interaction did not perturb the quadruplex folding of rG4. Together, these results establish G3BP1 as an rG4-recognition protein.

The C-Terminal RGG Domain of G3BP1 Is Involved with rG4 Recognition

G3BP1 is an evolutionarily conserved, multi-domain protein harboring an N-terminal NTF2-like domain, an acidic residue-rich region, PXXP motifs, an RNA-recognition motif (RRM) and a C-terminal arginine-glycine-glycine (RGG) domain (Figure 4.2D) (52,53). The RRM and RGG domains are well-known RNA-binding modules (54); therefore, we assessed whether these domains of G3BP1 are required for rG4 recognition.

We generated several truncated forms of G3BP1 protein, including those with the RRM and RGG domains being deleted (ΔRRM-RGG, 1-339 aa), or carrying both the RRM and RGG domains (340-466 aa, G3BP1-RRM-RGG) or either domain alone (340-415 aa, G3BP1-RRM and 430-466 aa, G3BP1-RGG). We next examined the binding capacities of these truncated forms of G3BP1 toward PITX1 rG4 using EMSA. We found that G3BP1-RRM and ΔRRM-RGG, both of

which lack the RGG domain, did not display appreciable interaction with rG4 (Figure 4.2E-F); G3BP1-RRM-RGG and G3BP1-RGG can bind with rG4, though the latter exhibited lower binding affinity than the former (Figure 4.2G-H).

We also measured the binding affinities of G3BP1-RRM-RGG toward rG4 and rM4 by using fluorescence anisotropy. The K_d values for the truncated G3BP1 in binding with PITX1 rG4 and rM4 were 78 ± 9 and 4500 ± 900 nM, respectively (Figure 4.2I); hence, G3BP1-RRM-RGG exhibits a comparable binding affinity and a much higher binding selectivity toward rG4 structure than the full-length protein. Together, these results demonstrated that the binding capacity of G3BP1 toward rG4 structure arises mainly from the C-terminal RGG domain and is enhanced by the RRM domain.

G3BP1 Regulates the Stability of PITX1 mRNA in an rG4-dependent Manner

Encouraged by the *in-vitro* binding results, we next examined the biological functions of G3BP1-rG4 interaction. One of the rG4 sequences employed for the aforementioned *in-vitro* binding assay was derived from the 3′-UTR of PITX1 mRNA, which was previously employed for assessing rG4-mediated regulation of mRNA translation (24,49). Thus, we chose PITX1 as a target to investigate whether G3BP1 can modulate the stability and translational efficiency of PITX1 mRNA through its binding with rG4 structures. To this end, we generated 293T cells with the *G3BP1* gene being knocked out with CRISPR-Cas9 (*G3BP1*$^{-/-}$). The successful depletion of G3BP1 protein in 293T cells was confirmed by Western blot (Figure 4.3A), and the deleted loci in genomic DNA were further identified by Sanger sequencing.

Results from Western blot analysis showed that the level of PITX1 protein was significantly increased after genetic depletion of G3BP1 in 293T cells (Figure 4.3B). This result suggests that G3BP1 may regulate the decay of PITX1 mRNA by binding to rG4 in its 3′-UTR or negatively modulate the translation of PITX1 mRNA. To distinguish these two scenarios, we

monitored the mRNA level of *PITX1* by RT-qPCR. The results showed that genetic depletion of G3BP1 in 293T cells led to a decrease in PITX1 mRNA level (Figure 4.3C), indicating that complete removal of G3BP1 may affect the stability of PITX1 mRNA. Together, G3BP1 positively modulates the stability of PITX1 mRNA, but it negatively regulates the translation of PITX1 mRNA.

We next investigated whether this regulation depends on G3BP1-rG4 interaction. Because PDS can disrupt the interactions between G3BP1 and rG4 structures *in vitro*, we examined how PDS treatment modulates the translation of PITX1 mRNA and whether this depends on G3BP1. Our results showed that PDS treatment led to an increased level of PITX1 protein in 293T cells (Figure 4.3D-F); this increase, however, was abolished in the isogenic *G3BP1*$^{-/-}$ cells (Figure 4.3G-I). Additionally, PDS treatment elicits a decrease in the mRNA level of *PITX1* in 293T cells, but not in *G3BP1*$^{-/-}$ cells (Figure 4.3F&I).

To further investigate the roles of G3BP1 and PDS in modulating mRNA stability, we measured the half-lives of PITX1 mRNA in 293T cells and *G3BP1*$^{-/-}$ cells with or without PDS treatment. As expected, genetic depletion of G3BP1 and PDS treatment both result in significant diminutions in the half-life of PITX1 mRNA in 293T cells (Figure 4.4A-B), with more pronounced effect being observed for PDS-treated cells. Additionally, PDS treatment can also diminish the half-life of PITX1 mRNA in *G3BP1*$^{-/-}$ cells (Figure 4.4A-B), indicating the potential role of PDS in regulating the stability of PITX1 mRNA through G3BP1-independent mechanism(s) (e.g., via displacing other proteins from rG4 structures in PITX1 mRNA). Together, these results underscore that PDS disrupts the interactions of G3BP1 with rG4 structures located in the 3′-UTR of PITX1 mRNA, thereby abrogating the regulatory effects of G3BP1 on the stability and translational efficiency of PITX1 mRNA.

To further determine whether the G3BP1-mediated regulation of PITX1 mRNA is rG4-dependent, we performed dual-luciferase reporter assay using the wild-type 3′-UTR of PITX1 mRNA (PITX1-WT) as well as its variant with all three rG4-forming sequences being mutated (PITX1-3Qm) (Figure 4.4C). We monitored both the mRNA and protein expression levels of firefly luciferase by using the corresponding signals from renilla luciferase as internal standard. Our results showed that genetic depletion of *G3BP1* led to markedly diminished level of luciferase mRNA arising from transcription of PITX1-WT plasmid **(**Figure 4.4D**)**. No appreciable difference in mRNA levels was, however, observed between two cell lines when transfected with PITX1-3Qm. These results support that G3BP1 positively regulates the stability of PITX1 mRNA and this regulation depends on rG4 structures located in the 3′-UTR of PITX1 mRNA. Moreover, the mRNA level of PITX1-3Qm was much lower than that of PITX1-WT in 293T cells, which again supports the role of the 3′-UTR rG4 structures in stabilizing of PITX1 mRNA.

In line with what was observed for endogenous PITX1 protein, we found that the firefly luciferase activity of PITX1-WT exhibited a pronounced increase upon genetic ablation of G3BP1 in 293T cells (Figure 4.4E). A similar increase was also observed for PITX1-3Qm (Figure 4.4E), indicating that G3BP1 also regulated the translation of PITX1 mRNA in an rG4-independent manner. A combination of mRNA and luciferase activity results showed that G3BP1 exerts a greater effect on the translational efficiency of PITX1-WT over PITX1-3Qm mRNA (Figure 4.4D-E). Additionally, the luciferase activity of PITX1-3Qm was also lower than PITX1-WT in *G3BP1*$^{-/-}$ cells (Figure 4.4E), underscoring that rG4s can also enhance the translation of PITX1 mRNA through a G3BP1-independent mechanism.

We also examined the expression levels of PITX1-WT and PITX1-3Qm in 293T cells with or without PDS treatment. We observed a significant decrease in the mRNA level of PITX1-WT upon PDS treatment (Figure 4.4D), which is in agreement with the aforementioned result of

the endogenous PITX1 mRNA (Figure 4.3I). PDS treatment, nevertheless, did not alter the mRNA level of PITX1-3Qm, and the treatment also abolished the difference in mRNA levels of PITX1-WT and PITX1-3Qm in 293T cells (Figure 4.4D). Moreover, while PDS treatment led to a slight decrease in the luciferase activity of PITX1-WT, the decrease is not as pronounced as that observed at the mRNA level (Figure 4.4D-E). This result suggests an elevated translational efficiency elicited by PDS treatment and parallels what we observed for endogenous PITX1 mRNA.

To explore further how G3BP1-rG4 interaction modulates the expression of PITX1 at mRNA and protein levels, we constructed Flag-PITX1-WT and Flag-PITX1-3Qm plasmids by replacing the coding sequence of the luciferase gene in the aforementioned PITX1-WT and PITX1-3Qm reporter plasmids with the coding sequence of a N-terminally Flag-tagged PITX1. Our RT-qPCR results revealed that the mRNA expression level of Flag-PITX1-WT was significantly higher than that of Flag-PITX1-3Qm in HEK293T cells ; genetic depletion of G3BP1, however, abolishes the difference observed for Flag-PITX1-WT and Flag-PITX1-3Qm. Our Western blot data showed that, upon transfection of HEK293T cells with the same amounts of plasmids, Flag-PITX1-3Qm protein was expressed at 45% level relative to that of Flag-PITX1-WT. At first glance, this appears to be incongruent with what we may predict from our observations made for the endogenous *PITX1* gene. Nevertheless, the Flag-PITX1 mRNA expressed from the transfected plasmid does not carry the 5′-UTR of endogenous PTIX1 mRNA, and it additionally contains the coding sequence for the Flag epitope tag. These differences may modulate differential recognitions of PITX1 mRNA by other RNA-binding proteins and/or ribosomes, thereby exerting different effects on the translation between endogenous PITX1 mRNA and that transcribed from the ectopically introduced plasmid. Nevertheless, genetic ablation of G3BP1 led to a markedly higher level (by > 20-fold) of expression of Flag-PITX1-

WT protein than Flag-PITX1-3Qm. Together, comparison of the above results obtained for Flag-PITX1-WT and Flag-PITX1-3Qm in HEK293T and the isogenic G3BP1-deficient background allowed us to conclude that G3BP1-rG4 interaction increases the stability of PITX1 mRNA and attenuates its translation efficiency, which mirrors what we observed for the endogenous PITX1 mRNA.

Together, these results support that G3BP1 positively regulates mRNA stability through its interaction with rG4 structures and negatively modulates the translation of mRNAs with rG4 structures in the 3′-UTRs. Nevertheless, our results also suggest the contributions of some mechanisms that are independent of rG4 and/or G3BP1 in these processes.

G3BP1 Binds with rG4 Structures in Cells to Regulate the Stabilities of Other mRNAs

To further explore the G3BP1-rG4 interactions in cells, we conducted seCLIP-seq experiments in HEK293T cells with ("PDS") or without ("Ctrl") PDS treatment. These experiments resulted in the identification of ~1400 and ~2900 mRNA binding sites for G3BP1 in "Ctrl" and "PDS" datasets, respectively (Figure 4.5A), and the overall signal intensities of G3BP1 peaks were higher in the "PDS" dataset than the "Ctrl" dataset (Figure 4.5B). The comparison between "Ctrl" and "PDS" seCLIP-seq datasets allowed for the identification of ~800 peaks with at least 1 bp of overlap (Figure 4.5A); among these overlapping peaks, ~100 contain putative rG4 sequences. Strikingly, the signal ratios of PDS/Ctrl of these rG4-containing peaks were significantly lower than the overall ratios of all overlapping peaks (Figure 4.5C), indicating the ability of PDS to displace G3BP1 from rG4 loci in the transcriptome. This finding corroborates the aforementioned *in-vitro* binding result and suggests that G3BP1 interacts with rG4 structures in cells.

The density plots obtained from metagene analysis showed that the binding sites of G3BP1 are enriched in the 5′-UTR, followed by 3′-UTR and CDS regions (Figure 4.5D).

However, after PDS treatment, the peak in the 3′-UTR completely shifted to the CDS regions, which is accompanied with a slight drop in signal in the 5′-UTR (Figure 4.5D), indicating that PDS can specifically disrupt the mRNA binding sites of G3BP1 at UTRs, especially the 3′-UTRs. Additionally, the peak annotation results showed an increased occupancy of G3BP1 in CDS regions, yet a decrease at the UTRs after PDS treatment.

We also compared the transcriptomic distributions of decreased peaks (ratios of PDS/Ctrl < 0.67, with at least 10 bp overlap) and increased peaks (ratios of PDS/Ctrl > 1.5, with at least 10 bp overlap) in "Ctrl" dataset after PDS treatment, as well as Ctrl-only peaks (no overlap with PDS). Here, we assumed that Ctrl-only peaks contain G3BP1 binding loci which are completely competed off by PDS. The density plots showed that Ctrl-only peaks and peaks that are decreased upon PDS treatment are predominately located in the UTRs, whereas those that are increased upon PDS treatment are mainly distributed in the CDS regions (Figure 4.5E). These results support that PDS displaces selectively G3BP1 from UTRs in mRNAs. Given that rG4 structures are highly enriched in UTRs (2), these results indicate the ability of PDS in disrupting specifically the interactions between G3BP1 and rG4 structure sites in cells. In this vein, it is worth noting that G3BP1 binding peaks around rG4 loci of *PITX1* and *NRAS* mRNAs were very weak, which might be due to the low abundance of these transcripts, dynamic nature of rG4 structures in these mRNAs, their dynamic interactions with G3BP1, and/or insufficient sequencing depth.

We next asked whether G3BP1 depletion and PDS treatment exert similar effects on the stabilities and translational efficiencies of target transcripts. We monitored the mRNA and protein levels of two representative genes (i.e. *KHSRP* and *ACTR2*), which were chosen on the basis of decreased G3BP1 occupancy at rG4 sites in UTRs after PDS treatment (Figure 4.5F**)**. The results showed that the mRNA levels of the *KHSRP* and *ACTR2* genes were markedly attenuated in G3BP1-depleted and PDS-treated cells, with the decreases being much more pronounced in PDS-

treated cells than G3BP1-depleted cells (Figure 4.5G-H). We also monitored the stabilities of KHSRP and ACTR2 mRNAs, and found that the half-lives of these two mRNAs were substantially decreased in *G3BP1*$^{-/-}$ and PDS-treated 293T cells relative to parental 293T cells without any treatment. Western blot results showed a significant increase in protein level of KHSRP in G3BP1-depleted cells, and an elevated translational efficiency of KHSRP upon PDS treatment (Figure 4.5I-J). Similarly, we observed a substantial increase in translational efficiency of ACTR2 mRNA in PDS-treated cells. In this regard, the ratio of expression levels of ACTR2 mRNA was 0.29 ± 0.06 in PDS-treated over untreated HEK293T cells, and the corresponding ratio of ACTR2 protein was 0.73 ± 0.14 (Figure 4.5I&K). Loss of G3BP1, on the other hand, led to a slight, yet statistically insignificant increase in translational efficiency of ACTR2 mRNA, where the ratios of mRNA and protein expression levels were 0.68 ± 0.13 and 0.81 ± 0.21, respectively, in *G3BP1*$^{-/-}$ over parental HEK293T cells (Figure 4.5I&J). The more pronounced effect exerted by PDS treatment over genetic ablation of *G3BP1* may be due to the translational regulation of G3BP1 mRNA by other protein(s) that can recognize rG4 structures in its 3′-UTR. Cumulatively, these results again support a role of G3BP1-rG4 interaction in modulating the stabilities and translational efficiencies of mRNAs.

We next examined whether the above findings made for 293T cells are general. To this end, we first investigated the effects of G3BP1 ablation and PDS treatment on the mRNA and protein levels of *PITX1*, *KHSRP* and *ACTR2* genes in HeLa cells. RT-qPCR and Western blot results showed that the mRNA levels of the three genes dropped pronouncedly after PDS treatment, which is accompanied with slight increases in translation efficiencies. In addition, CRISPR-mediated genetic ablation of *G3BP1* in HeLa cells led to diminished mRNA levels and elevated translation efficiencies of these three genes. Moreover, by analyzing the publicly available RNA-seq data from ENCODE, we found decreased mRNA levels of the three genes in

G3BP1-depleted HepG2 and K562 cells. These results suggest that the effects of G3BP1 depletion on rG4-bearing transcripts are general in mammalian cells.

4.2. Conclusions

G-rich sequences in RNAs can fold into rG4 structures, which modulate the stabilities and translational efficiencies of mRNAs (55). Some RBPs, which can bind to rG4s in their folded and unfolded forms, are highly correlated with the rG4-mediated regulations of mRNAs. For instance, RNA helicases in cells can bind to and unwind rG4 structures into single-stranded RNAs which are subsequently recognized by G-rich element-binding proteins (e.g. hnRNP H/F and CNBP) and prevent their refolding into rG4 structures, thereby increasing the translational efficiencies of mRNAs (26,56). In addition, GRSF1 regulates the degradation of rG4-containing mRNAs through unwinding rG4 structures to facilitate degradosome-mediated decay (57). Therefore, there is a growing interest in identifying rG4-binding proteins and characterizing their functions.

We employed a bioinformatic approach, relying on the analysis of overlapping peaks between publicly available rG4-seq dataset (2) and eCLIP-seq datasets for RBPs (28), to uncover putative novel rG4-binding proteins. Our analysis captured some previously reported rG4-binding proteins, which display high levels of occupancy at rG4 loci in the transcriptome (Figure 4.1), validating the ability of the method in identifying candidate rG4-binding proteins. Importantly, our analysis also led to the identification of a very large number of putative rG4-binding proteins whose capabilities in binding to rG4 structures have not been previously documented. This provides an important list of proteins for the research community to assess their direct interactions with rG4s and to explore their biological functions. It is worth noting a limitation of our analysis. In particular, the rG4-seq dataset and eCLIP-seq datasets were generated from

different cell lines and the distributions in rG4 in mRNAs and RNA-protein interactions can vary with cell lines, which may lead to false-negative discovery of putative rG4-binding proteins.

Among the top-ranked candidate rG4-binding proteins are splicing factors (e.g. FMR1, RMB15, PPIG, PRPF8 and SRSF1) or helicases (e.g. AQR, DDX3X, DDX55, DDX6, G3BP1 and UPF1). These results are in accordance with the fact that rG4s are implicated in control of mRNA processing and translation through splicing factors and RNA helicases, respectively (11,14,49).

G3BP1 plays an essential role in stress granule formation, DNA-triggered cGAS/STING pathway, RIG-I-mediated cellular antiviral response and innate immune response (58-61). It also displays Mg^{2+}- and ATP-dependent helicase activity (62). Here, we identified G3BP1 as a direct rG4-binding protein with low-nM binding affinities, which are much stronger than those for binding with the corresponding rM4 probes (Figure 4.2A-C). We also found that the C-terminal RGG domain of G3BP1 is indispensable for its binding toward rG4 structures (Figure 4.2E-H). Moreover, the G3BP1-RRM-RGG truncated protein exhibits a much higher rG4-binding selectivity than the full-length G3BP1 protein (Figure 4.2I). In this vein, RGG domain is the second most common RNA-binding domain present in the human proteome (63), and several known G4-binding proteins, including CIRBP, FMRP and TLS/FUS, recognize G4 structures through their RGG domains (45,64,65).

We also found that PDS, a small-molecule G4 ligand, could disrupt pronouncedly the G3BP1-rG4 interaction *in vitro*. In addition, our seCLIP-seq results revealed that, upon PDS treatment, the signal intensities of rG4-containing peaks decreased (Figure 4.5C) and the transcriptome-wide distribution of G3BP1-binding sites exhibited a drastic shift from 3′-UTRs to CDS regions (Figure 4.5D-E**)**. Given that rG4-forming sequences are highly enriched in UTRs

(2), our results provide strong evidence to support that PDS can disrupt G3BP1-rG4 interactions in cells.

Over the last few years, increasing lines of evidence support that rG4 structures assume critical roles in regulating pre-mRNA processing (splicing and polyadenylation), mRNA stability and translation (1,17,24). These regulatory processes often entail rG4-binding proteins to modulate G4 conformation and/or serve as bridges to recruit additional regulatory proteins. Here we demonstrated that G3BP1 can enhance the stability and suppress the translational efficiency of PITX1 mRNA, which harbors three rG4 structures in its 3′-UTR. Further analysis revealed that treatment with PDS and genetic depletion of G3BP1 (Figure 4.3), both of which disrupted the interactions between G3BP1 and rG4 structures in PITX1 mRNA, abolished the regulatory effect of G3BP1, underscoring the importance of G3BP1-rG4 interaction in RNA metabolism. This notion finds additional support from dual-luciferase reporter assay results, showing that loss of G3BP1 and PDS treatment led to diminished mRNA levels and elevated translation of luciferase mRNA harboring PITX1 3′-UTR (PITX1-WT). The modulatory effects of G3BP1 depletion and PDS treatment were also observed for two other target transcripts (i.e. *KHSRP* and *ACTR2*) chosen based on seCLIP-seq results (Figure 4.5G-H), further illustrating the rG4-dependent regulatory functions of G3BP1 on the stabilities and translation efficiencies of mRNA. In this vein, the roles of rG4 in modulating the stabilities of mRNAs also find support from the observation that genetic depletion of some putative rG4-binding proteins led to more pronounced alterations in mRNA expressions of those genes with putative G4 structures in the UTRs than those without (66).

G3BP1 is the central node and molecular switch that trigger RNA-protein phase separation (67,68), and rG4 structures can promote RNA phase separation (69,70). Our findings are in agreement with the previous observation that the formation of G3BP1-mRNA

ribonucleoprotein particles or stress granules could protect mRNAs from degradation, while concomitantly confer a poor translation efficiency (60). It is of note that this G3BP1-mediated effect was observed in stressed cells, and there is little evidence to support a similar function of G3BP1 in unstressed cells. Another possibility is that, G3BP1 preferentially interacts with G4-containing RNAs, which may attenuate partially the auto-inhibitory effect of G3BP1 under normal conditions (67), thereby increasing the accessibility of G3BP1 to other RNAs and proteins to form large protein-RNA complexes. Like ribonucleoprotein particles and stress granules, these protein-RNA complexes can protect mRNAs from degradation and suppress their translation. Thus, it is also possible that G3BP1-rG4 complex acts as a scaffold to recruit other RNAs and proteins.

G3BP1 contains an NTF2-like domain and multiple intrinsically disordered regions (IDRs), which regulate the dimerization of G3BP1 and G3BP1-RNA interaction, respectively (67). The IDRs of G3BP1 resemble the RNA recognition motifs (RRMs) of other stress granule proteins, including hnRNPA1 and hnRNPA2B1, whose RRMs were shown to be capable of substituting the IDRs of G3BP1 to support the RNA-dependent liquid-liquid phase separation and stress granules assembly (67). Different from other stress granule proteins, G3BP1 exhibits intramolecular interaction between IDR1 and IDR3, and is susceptible to dimerization via the NTF2-like domain, which are essential for liquid-liquid phase separation and the maintenance of stress granules in cells. Hence, these two properties may endow G3BP1 the ability to be the core component of the stress granule network.

It will be important to examine, in the future, other top-ranked candidate rG4-binding proteins identified from our bioinformatics analysis, e.g., UPF1, DDX55, DDX6 and RBM15. In this regard, UPF1, an ATP-dependent RNA helicase, was found to regulate the decay of highly structured RNA in cooperation with G3BP1 (48) and target GC-rich region to trigger RNA decay

(71). Notably, the function of G3BP1 in the UPF1-G3BP1-mediated RNA decay is distinct from its protective role proposed here. In the UPF1-G3BP1-mediated RNA decay, UPF1 dominates the regulation by recognizing highly structured RNAs especially double-stranded RNAs, unwinding them, thereby facilitating the enrichment of G3BP1 in proximity to UPF1 (48). Thus, the differences in modes of interactions between G3BP1 and targeted transcripts may contribute to the different functions of G3BP1 in modulating mRNA stabilities.

In summary, we identified multiple candidate rG4-binding proteins with a bioinformatic approach, and we validated that one of these proteins, G3BP1 can bind directly with rG4 with low-nM binding affinity. We also found that the binding of G3BP1 with rG4 structures in the UTRs stabilized mRNAs and suppressed their translation, which revealed a new function of G3BP1. Together, the results from the present study uncovered a number of candidate rG4-binding proteins and expanded the functions of G3BP1.

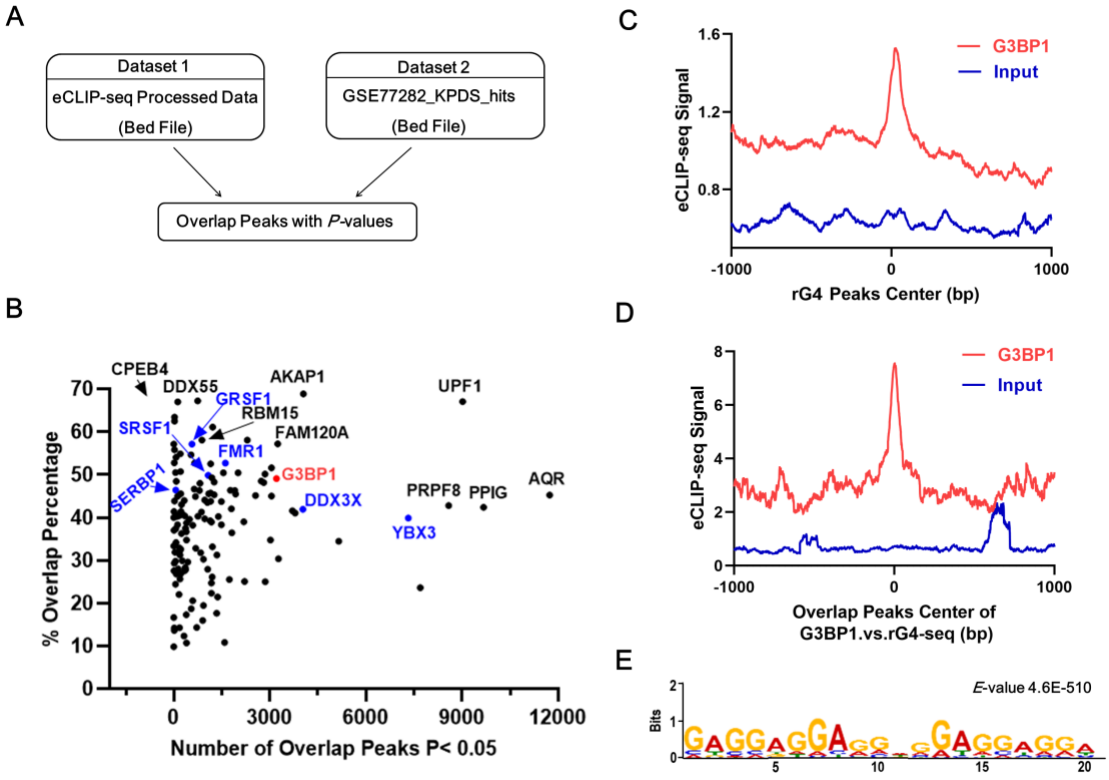

Figure 4.1. Bioinformatic discovery of rG4-binding proteins. (A) A schematic diagram showing the workflow of the bioinformatic method. (B) A scatter plot illustrating the results obtained from bioinformatic analysis. (C) The distribution of G3BP1 eCLIP-seq peaks relative to the center of rG4-seq peaks; (D) the center of the overlapping peaks between G3BP1 eCLIP-seq and rG4-seq datasets. "Input" represents size-match input dataset (ENCSR907EBB), which is a control eCLIP-seq dataset. (E) The most enriched motif of the overlapping peaks.

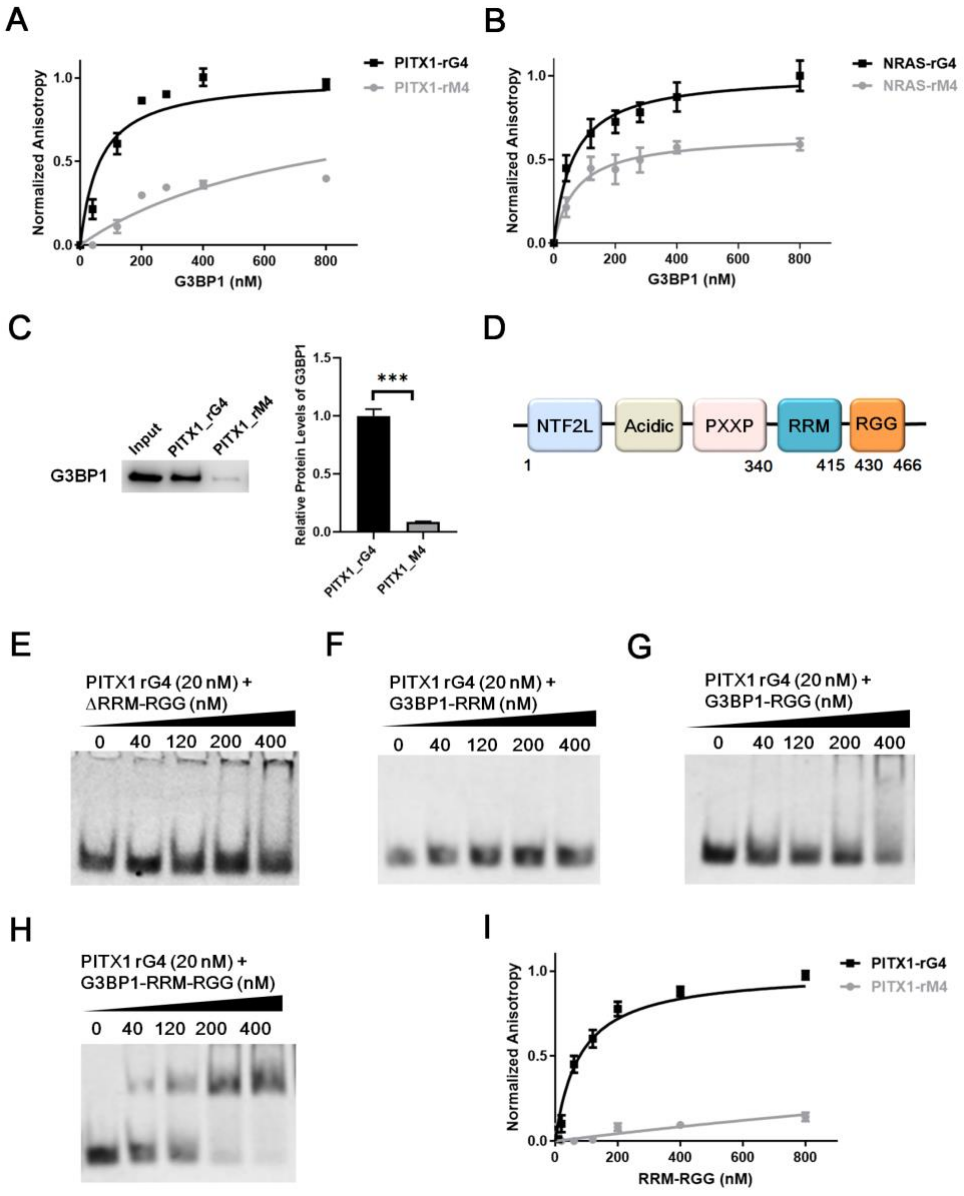

Figure 4.2. G3BP1 binds directly and selectively with rG4 structures. Fluorescence anisotropy for measuring the binding affinities of G3BP1 protein toward rG4 structures derived from PITX1 (A) and NRAS (B) mRNAs and their corresponding mutants (rM4s). Error bars bars represent S.D. (n = 3). (C) Western blot images and quantitative results obtained from in vitro pull-down of G3BP1 protein from whole-cell protein lysates with the use of biotin-labeled PITX1 rG4 and rM4 probes. Error bars represent S.E.M. (n = 3). (D) A schematic diagram depicting the domain structure of G3BP1 protein. (E-H) EMSA for monitoring the interactions between the four truncated forms of G3BP1 with PITX1 rG4 probe. (I) Fluorescence anisotropy for measuring the binding affinities of G3BP1-RRM-RGG with PITX1 rG4 and rM4. Error bars represent S.D. (n = 3). The p values were calculated by unpaired, two-tailed Student's t-test. ***, $p < 0.001$.

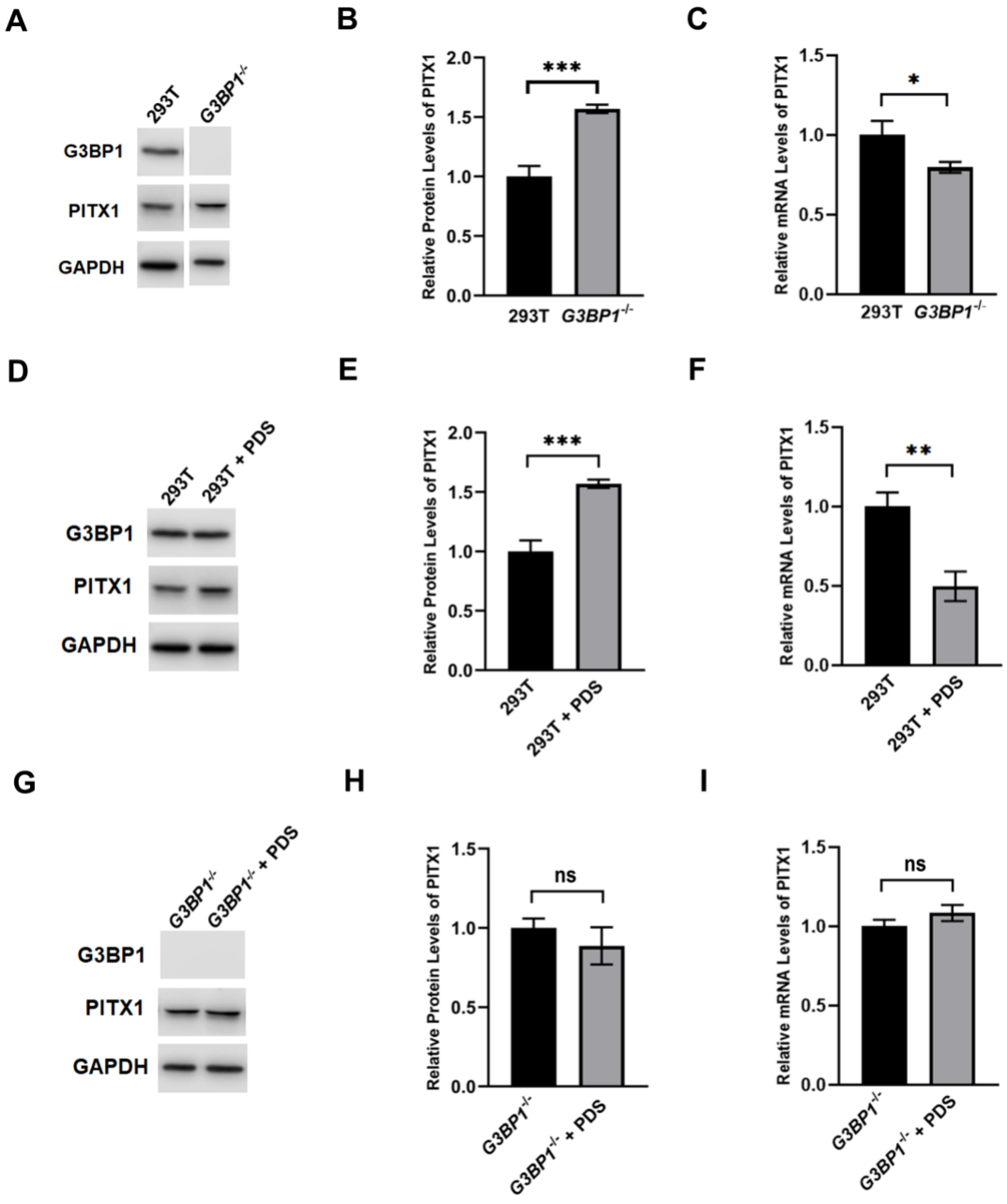

Figure 4.3. Genetic depletion of G3BP1 and PDS treatment confer similar effects on up-regulating the protein level and down-regulating the mRNA level of PITX1. Western blot and RT-qPCR analyses for monitoring the protein and mRNA levels of PITX1 in 293T cells and the isogenic G3BP1-/- cells (A-C), in 293T cells with or without PDS treatment (D-F), and in G3BP1-/- cells with or without PDS treatment (G-I). Error bars represent S.D. (n = 3). The p values were calculated by unpaired, two-tailed Student's t-test. *, $0.01 < p < 0.05$; **, $0.001 < p < 0.01$; ***, $p < 0.001$.

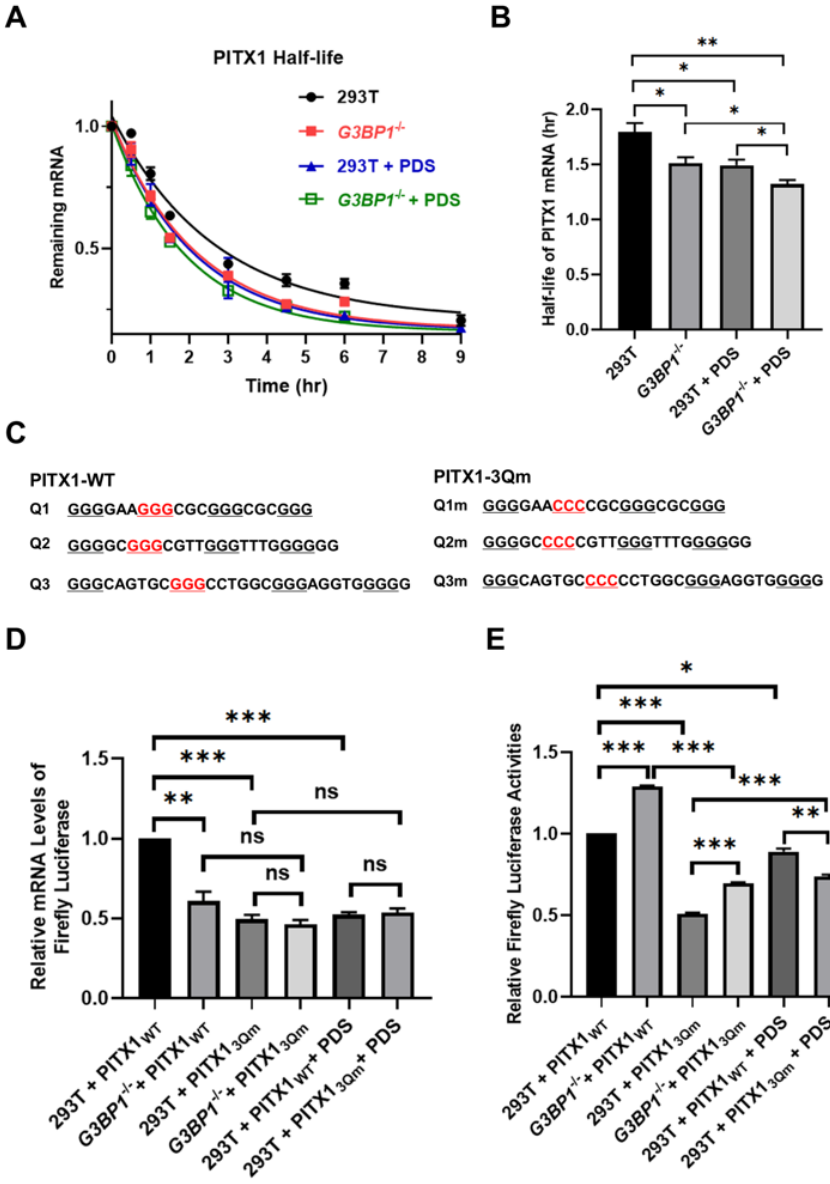

Figure 4.4. G3BP1 regulates the stability of PITX1 mRNA through its interaction with rG4 structures. (A) RT-qPCR results showing the half-lives of PITX1 mRNA in 293T cells, G3BP1-/- cells, PDS-treated 293T cells, and PDS-treated G3BP1-/- cells. (B) Bar chart showing the half-lives calculated from the above RT-qPCR results. (C) rG4-forming sequences in the 3'-UTR of PITX1 mRNA and their corresponding mutants. (D) Relative firefly luciferase mRNA levels (normalized to the level of renilla luciferase mRNA) and (E) relative firefly luciferase activities (normalized to renilla luciferase activity) in 293T cells, G3BP1-/- cells, and PDS-treated 293T cells expressed from PITX1-WT or PITX1-3Qm plasmid. Error bars represent S.E.M. (n = 3). The p values were calculated by using unpaired, two-tailed Student's *t*-test. ns, $p < 0.05$; **, $0.001 < p < 0.01$; ***, $p < 0.001$.

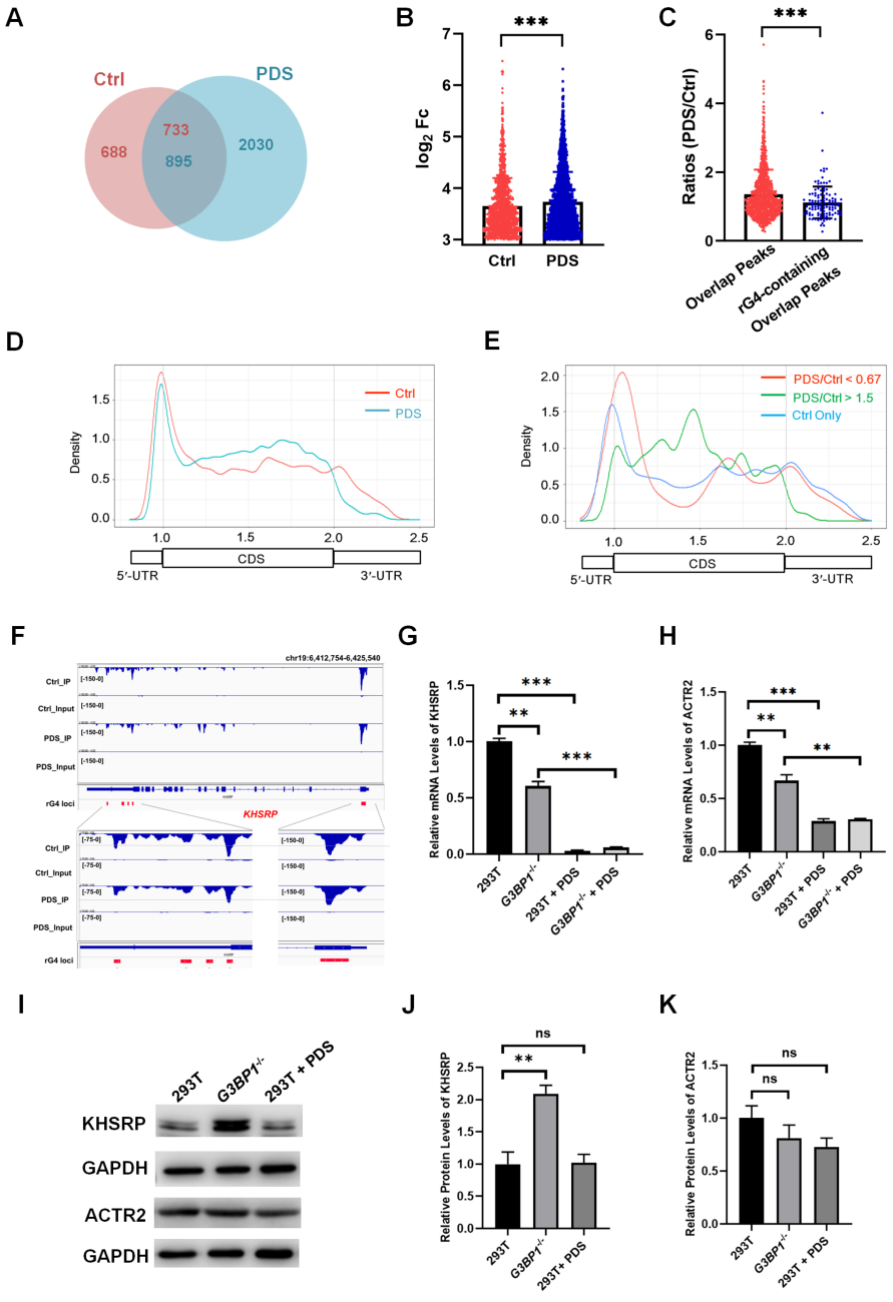

Figure 4.5. seCLIP-seq analyses of G3BP1 in 293T cells without ("Ctrl") or with ("PDS") PDS treatment reveal the G3BP1-rG4 interaction in cells. (A) A Venn diagram showing the overlap between "Ctrl" and "PDS" datasets. (B) The comparison between "Ctrl" and "PDS" peak intensities in $\log_2$(FoldChange) of IP sample and Input sample. (C) Signal ratios of PDS/Ctrl in all overlapping peaks and rG4-containing overlapping peaks. (D) Metagene analyses for profiling the transcriptomic distributions of "Ctrl" and "PDS" datasets. (E) Metagene analyses for profiling the transcriptomic distributions of decreased peaks, increased peaks and Ctrl-only peaks

in “Ctrl” dataset relative to the “PDS” dataset. (F) IGV plots showing the comparison of “Ctrl” and “PDS” peaks around the G4-forming sequences located in the 5’- and 3’-UTRs of KHSRP mRNA. “Input” represents size-match input sample. (G-H) RT-qPCR results showing the relative mRNA levels of KHSRP (G) and ACTR2 (H) genes in 293T cells and G3BP1-/- cells without or with PDS treatment. (I-K) Western blot analysis for monitoring the protein levels of KHSRP (J) and ACTR2 (K) genes in 293T cells and G3BP1-/- cells, and PDS-treated 293T cells. Error bars represent S.D. (n = 3). The p values were calculated by using unpaired, two-tailed Student’s t-test. ns, $p < 0.05$; **, $0.001 < p < 0.01$; ***, $p < 0.001$.

4.3. References

1. Millevoi, S., Moine, H. and Vagner, S. (2012) G-quadruplexes in RNA biology. *Wiley Interdiscip Rev: RNA*, **3**, 495-507.

2. Kwok, C.K., Marsico, G., Sahakyan, A.B., Chambers, V.S. and Balasubramanian, S. (2016) rG4-seq reveals widespread formation of G-quadruplex structures in the human transcriptome. *Nat Methods*, **13**, 841.

3. Guo, J.U. and Bartel, D.P. (2016) RNA G-quadruplexes are globally unfolded in eukaryotic cells and depleted in bacteria. *Science (New York, N.Y.)*, **353**, aaf5371.

4. Puig Lombardi, E. and Londoño-Vallejo, A. (2019) A guide to computational methods for G-quadruplex prediction. *Nucleic Acids Res*, **48**, 1-15.

5. Yang, S.Y., Lejault, P., Chevrier, S., Boidot, R., Robertson, A.G., Wong, J.M.Y. and Monchaud, D. (2018) Transcriptome-wide identification of transient RNA G-quadruplexes in human cells. *Nat Commun*, **9**, 4730.

6. Biffi, G., Di Antonio, M., Tannahill, D. and Balasubramanian, S. (2013) Visualization and selective chemical targeting of RNA G-quadruplex structures in the cytoplasm of human cells. *Nat. Chem.*, **6**, 75.

7. Laguerre, A., Hukezalie, K., Winckler, P., Katranji, F., Chanteloup, G., Pirrotta, M., Perrier-Cornet, J.-M., Wong, J.M.Y. and Monchaud, D. (2015) Visualization of RNA-Quadruplexes in Live Cells. *J Am Chem Soc*, **137**, 8521-8525.

8. Chen, X.-C., Chen, S.-B., Dai, J., Yuan, J.-H., Ou, T.-M., Huang, Z.-S. and Tan, J.-H. (2018) Tracking the Dynamic Folding and Unfolding of RNA G-Quadruplexes in Live Cells. *Angew Chem Int Edit*, **57**, 4702-4706.

9. Weng, X., Gong, J., Chen, Y., Wu, T., Wang, F., Yang, S., Yuan, Y., Luo, G., Chen, K., Hu, L. *et al.* (2020) Keth-seq for transcriptome-wide RNA structure mapping. *Nat Chem Biol*, **16**, 489-492.

10. Huppert, J.L., Bugaut, A., Kumari, S. and Balasubramanian, S. (2008) G-quadruplexes: the beginning and end of UTRs. *Nucleic Acids Res*, **36**, 6260-6268.

11. Wolfe, A.L., Singh, K., Zhong, Y., Drewe, P., Rajasekhar, V.K., Sanghvi, V.R., Mavrakis, K.J., Jiang, M., Roderick, J.E., Van der Meulen, J. *et al.* (2014) RNA G-quadruplexes cause eIF4A-dependent oncogene translation in cancer. *Nature*, **513**, 65.

12. Darnell, J.C., Jensen, K.B., Jin, P., Brown, V., Warren, S.T. and Darnell, R.B. (2001) Fragile X Mental Retardation Protein Targets G Quartet mRNAs Important for Neuronal Function. *Cell*, **107**, 489-499.

13. Kumari, S., Bugaut, A., Huppert, J.L. and Balasubramanian, S. (2007) An RNA G-quadruplex in the 5' UTR of the NRAS proto-oncogene modulates translation. *Nat Chem Biol*, **3**, 218-221.

14. Cammas, A. and Millevoi, S. (2017) RNA G-quadruplexes: emerging mechanisms in disease. *Nucleic Acids Res*, **45**, 1584-1595.

15. Fay, M.M., Lyons, S.M. and Ivanov, P. (2017) RNA G-Quadruplexes in Biology: Principles and Molecular Mechanisms. *J Mol Biol*, **429**, 2127-2147.

16. Kwok, C.K. and Merrick, C.J. (2017) G-Quadruplexes: Prediction, Characterization, and Biological Application. *Trends Biotechnol*, **35**, 997-1013.

17. Sauer, M., Juranek, S.A., Marks, J., De Magis, A., Kazemier, H.G., Hilbig, D., Benhalevy, D., Wang, X., Hafner, M. and Paeschke, K. (2019) DHX36 prevents the accumulation of translationally inactive mRNAs with G4-structures in untranslated regions. *Nat Commun*, **10**, 2421.

18. Herdy, B., Mayer, C., Varshney, D., Marsico, G., Murat, P., Taylor, C., D'Santos, C., Tannahill, D. and Balasubramanian, S. (2018) Analysis of NRAS RNA G-quadruplex binding proteins reveals DDX3X as a novel interactor of cellular G-quadruplex containing transcripts. *Nucleic Acids Res*, **46**, 11592-11604.

19. Serikawa, T., Spanos, C., von Hacht, A., Budisa, N., Rappsilber, J. and Kurreck, J. (2018) Comprehensive identification of proteins binding to RNA G-quadruplex motifs in the 5′ UTR of tumor-associated mRNAs. *Biochimie*, **144**, 169-184.

20. von Hacht, A., Seifert, O., Menger, M., Schütze, T., Arora, A., Konthur, Z., Neubauer, P., Wagner, A., Weise, C. and Kurreck, J. (2014) Identification and characterization of RNA guanine-quadruplex binding proteins. *Nucleic Acids Res*, **42**, 6630-6644.

21. Bian, W.-X., Xie, Y., Wang, X.-N., Xu, G.-H., Fu, B.-S., Li, S., Long, G., Zhou, X. and Zhang, X.-L. (2018) Binding of cellular nucleolin with the viral core RNA G-quadruplex structure suppresses HCV replication. *Nucleic Acids Res*, **47**, 56-68.

22. Lista, M.J., Martins, R.P., Billant, O., Contesse, M.-A., Findakly, S., Pochard, P., Daskalogianni, C., Beauvineau, C., Guetta, C., Jamin, C. *et al.* (2017) Nucleolin directly mediates Epstein-Barr virus immune evasion through binding to G-quadruplexes of EBNA1 mRNA. *Nat Commun*, **8**, 16043.

23. Sexton, A.N. and Collins, K. (2011) The 5′ Guanosine Tracts of Human Telomerase RNA Are Recognized by the G-Quadruplex Binding Domain of the RNA Helicase DHX36 and Function To Increase RNA Accumulation. *Mol Cell Biol*, **31**, 736-743.

24. McRae, E.K.S., Booy, E.P., Moya-Torres, A., Ezzati, P., Stetefeld, J. and McKenna, S.A. (2017) Human DDX21 binds and unwinds RNA guanine quadruplexes. *Nucleic Acids Res*, **45**, 6656-6668.

25. Zhang, Y., Gaetano, C.M., Williams, K.R., Bassell, G.J. and Mihailescu, M.R. (2014) FMRP interacts with G-quadruplex structures in the 3'-UTR of its dendritic target Shank1 mRNA. *RNA Biology*, **11**, 1364-1374.

26. Benhalevy, D., Gupta, S.K., Danan, C.H., Ghosal, S., Sun, H.-W., Kazemier, H.G., Paeschke, K., Hafner, M. and Juranek, S.A. (2017) The Human CCHC-type Zinc Finger Nucleic Acid-Binding Protein Binds G-Rich Elements in Target mRNA Coding Sequences and Promotes Translation. *Cell Rep*, **18**, 2979-2990.

27. Thandapani, P., Song, J., Gandin, V., Cai, Y., Rouleau, S.G., Garant, J.M., Boisvert, F.M., Yu, Z., Perreault, J.P., Topisirovic, I. *et al.* (2015) Aven recognition of RNA G-quadruplexes regulates translation of the mixed lineage leukemia protooncogenes. *Elife*, **4**, 1-30.

28. Van Nostrand, E.L., Pratt, G.A., Shishkin, A.A., Gelboin-Burkhart, C., Fang, M.Y., Sundararaman, B., Blue, S.M., Nguyen, T.B., Surka, C., Elkins, K. *et al.* (2016) Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat Methods*, **13**, 508.

29. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57-74.

30. Consortium, E.P. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57-74.

31. Chikina, M.D. and Troyanskaya, O.G. (2012) An effective statistical evaluation of ChIPseq dataset similarity. *Bioinformatics*, **28**, 607-613.

32. Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H. and 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**, 576-589.

33. Pohl, A. and Beato, M. (2014) bwtool: a tool for bigWig files. *Bioinformatics*, **30**, 1618-1619.

34. Thorvaldsdóttir, H., Robinson, J.T. and Mesirov, J.P. (2012) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings Bioinf*, **14**, 178-192.

35. Williams, P., Li, L., Dong, X. and Wang, Y. (2017) Identification of SLIRP as a G quadruplex-binding protein. *Journal of the American Chemical Society*, **139**, 12426-12429.

36. Hellman, L.M. and Fried, M.G. (2007) Electrophoretic mobility shift assay (EMSA) for detecting protein–nucleic acid interactions. *Nat Protoc*, **2**, 1849.

37. Ran, F.A., Hsu, P.D., Wright, J., Agarwala, V., Scott, D.A. and Zhang, F. (2013) Genome engineering using the CRISPR-Cas9 system. *Nat Protoc*, **8**, 2281.

38. Doench, J.G., Fusi, N., Sullender, M., Hegde, M., Vaimberg, E.W., Donovan, K.F., Smith, I., Tothova, Z., Wilen, C., Orchard, R. *et al.* (2016) Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat Biotechnol*, **34**, 184-191.

39. Liu, H. and Naismith, J.H. (2008) An efficient one-step site-directed deletion, insertion, single and multiple-site plasmid mutagenesis protocol. *BMC Biotechnol*, **8**, 91-91.

40. Van Nostrand, E.L., Nguyen, T.B., Gelboin-Burkhart, C., Wang, R., Blue, S.M., Pratt, G.A., Louie, A.L. and Yeo, G.W. (2017) Robust, Cost-Effective Profiling of RNA Binding Protein Targets with Single-end Enhanced Crosslinking and Immunoprecipitation (seCLIP). *Methods Mol Biol*, **1648**, 177-200.

41. Quinlan, A.R. and Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841-842.

42. Olarerin-George, A.O. and Jaffrey, S.R. (2017) MetaPlotR: a Perl/R pipeline for plotting metagenes of nucleotide modifications and other transcriptomic sites. *Bioinformatics (Oxford, England)*, **33**, 1563-1564.

43. Bedrat, A., Lacroix, L. and Mergny, J.-L. (2016) Re-evaluation of G-quadruplex propensity with G4Hunter. *Nucleic Acids Res*, **44**, 1746-1759.

44. Chikina, M.D. and Troyanskaya, O.G. (2012) An effective statistical evaluation of ChIPseq dataset similarity. *Bioinformatics (Oxford, England)*, **28**, 607-613.

45. Phan, A.T., Kuryavyi, V., Darnell, J.C., Serganov, A., Majumdar, A., Ilin, S., Raslin, T., Polonskaia, A., Chen, C., Clain, D. *et al.* (2011) Structure-function studies of FMRP RGG peptide recognition of an RNA duplex-quadruplex junction. *Nat Struct Mol Biol*, **18**, 796-804.

46. Brázda, V., Kolomazník, J., Lýsek, J., Bartas, M., Fojta, M., Šťastný, J. and Mergny, J.-L. (2019) G4Hunter web application: a web server for G-quadruplex prediction. *Bioinformatics*, **35**, 3493-3495.

47. Beaudoin, J.-D., Jodoin, R. and Perreault, J.-P. (2014) New scoring system to identify RNA G-quadruplex folding. *Nucleic Acids Res*, **42**, 1209-1223.

48. Fischer, J.W., Busa, V.F., Shao, Y. and Leung, A.K.L. (2020) Structure-Mediated RNA Decay by UPF1 and G3BP1. *Mol Cell*, **78**, 70-84.e76.

49. Booy, E.P., Howard, R., Marushchak, O., Ariyo, E.O., Meier, M., Novakowski, S.K., Deo, S.R., Dzananovic, E., Stetefeld, J. and McKenna, S.A. (2014) The RNA helicase RHAU (DHX36) suppresses expression of the transcription factor PITX1. *Nucleic Acids Res*, **42**, 3346-3361.

50. Paramasivan, S., Rujan, I. and Bolton, P.H. (2007) Circular dichroism of quadruplex DNAs: Applications to structure, cation effects and ligand binding. *Methods*, **43**, 324-331.

51. Rodriguez, R., Müller, S., Yeoman, J.A., Trentesaux, C., Riou, J.-F. and 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.

52. Tourrière, H., Chebli, K., Zekri, L., Courselaud, B., Blanchard, J.M., Bertrand, E. and Tazi, J. (2003) The RasGAP-associated endoribonuclease G3BP assembles stress granules. *J Cell Biol*, **160**, 823-831.

53. Parker, F., Maurier, F., Delumeau, I., Duchesne, M., Faucher, D., Debussche, L., Dugue, A., Schweighoffer, F. and Tocque, B. (1996) A Ras-GTPase-activating protein SH3-domain-binding protein. *Mol Cell Biol*, **16**, 2561-2569.

54. Castello, A., Fischer, B., Frese, Christian K., Horos, R., Alleaume, A.-M., Foehr, S., Curk, T., Krijgsveld, J. and Hentze, Matthias W. (2016) Comprehensive Identification of RNA-Binding Domains in Human Cells. *Mol Cell*, **63**, 696-710.

55. Dumas, L., Herviou, P., Dassi, E., Cammas, A. and Millevoi, S. (2020) G-Quadruplexes in RNA Biology: Recent Advances and Future Directions. *Trends Biochem Sci*, **46**, 270-283.

56. Dominguez, C., Fisette, J.-F., Chabot, B. and Allain, F.H.T. (2010) Structural basis of G-tract recognition and encaging by hnRNP F quasi-RRMs. *Nat Struct Mol Biol*, **17**, 853-861.

57. Pietras, Z., Wojcik, M.A., Borowski, L.S., Szewczyk, M., Kulinski, T.M., Cysewski, D., Stepien, P.P., Dziembowski, A. and Szczesny, R.J. (2018) Dedicated surveillance mechanism controls G-quadruplex forming non-coding RNAs in human mitochondria. *Nat Commun*, **9**, 2558.

58. Markmiller, S., Soltanieh, S., Server, K.L., Mak, R., Jin, W., Fang, M.Y., Luo, E.-C., Krach, F., Yang, D., Sen, A. *et al.* (2018) Context-Dependent and Disease-Specific Diversity in Protein Interactions within Stress Granules. *Cell*, **172**, 590-604.e513.

59. Onomoto, K., Yoneyama, M., Fung, G., Kato, H. and Fujita, T. (2014) Antiviral innate immunity and stress granule responses. *Trends Immunol*, **35**, 420-428.

60. Khong, A., Matheny, T., Jain, S., Mitchell, S.F., Wheeler, J.R. and Parker, R. (2017) The Stress Granule Transcriptome Reveals Principles of mRNA Accumulation in Stress Granules. *Mol Cell*, **68**, 808-820.e805.

61. Liu, Z.-S., Cai, H., Xue, W., Wang, M., Xia, T., Li, W.-J., Xing, J.-Q., Zhao, M., Huang, Y.-J., Chen, S. *et al.* (2019) G3BP1 promotes DNA binding and activation of cGAS. *Nat Immunol*, **20**, 18-28.

62. Costa, M., Ochem, A., Staub, A. and Falaschi, A. (1999) Human DNA helicase VIII: a DNA and RNA helicase corresponding to the G3BP protein, an element of the ras transduction pathway. *Nucleic Acids Res*, **27**, 817-821.

63. Thandapani, P., O'Connor, T.R., Bailey, Timothy L. and Richard, S. (2013) Defining the RGG/RG Motif. *Mol Cell*, **50**, 613-623.

64. Huang, Z.-L., Dai, J., Luo, W.-H., Wang, X.-G., Tan, J.-H., Chen, S.-B. and Huang, Z.-S. (2018) Identification of G-Quadruplex-Binding Protein from the Exploration of RGG Motif/G-Quadruplex Interactions. *J Am Chem Soc*, **140**, 17945-17955.

65. Takahama, K., Takada, A., Tada, S., Shimizu, M., Sayama, K., Kurokawa, R. and Oyoshi, T. (2013) Regulation of Telomere Length by G-Quadruplex Telomere DNA- and TERRA-Binding Protein TLS/FUS. *Chem Biol*, **20**, 341-350.

66. Lee, D.S.M., Ghanem, L.R. and Barash, Y. (2020) Integrative analysis reveals RNA G-quadruplexes in UTRs are selectively constrained and enriched for functional associations. *Nat Commun*, **11**, 527-527.

67. Yang, P., Mathieu, C., Kolaitis, R.-M., Zhang, P., Messing, J., Yurtsever, U., Yang, Z., Wu, J., Li, Y., Pan, Q. *et al.* (2020) G3BP1 Is a Tunable Switch that Triggers Phase Separation to Assemble Stress Granules. *Cell*, **181**, 325-345.e328.

68. Guillén-Boixet, J., Kopach, A., Holehouse, A.S., Wittmann, S., Jahnel, M., Schlüßler, R., Kim, K., Trussina, I.R.E.A., Wang, J., Mateju, D. *et al.* (2020) RNA-Induced Conformational Switching and Clustering of G3BP Drive Stress Granule Assembly by Condensation. *Cell*, **181**, 346-361.e317.

69. Jain, A. and Vale, R.D. (2017) RNA phase transitions in repeat expansion disorders. *Nature*, **546**, 243.

70. Zhang, Y., Yang, M., Duncan, S., Yang, X., Abdelhamid, M.A.S., Huang, L., Zhang, H., Benfey, P.N., Waller, Z.A.E. and Ding, Y. (2019) G-quadruplex structures trigger RNA phase separation. *Nucleic Acids Res*, **47**, 11746-11754.

71. Imamachi, N., Salam, K.A., Suzuki, Y. and Akimitsu, N. (2017) A GC-rich sequence feature in the 3′ UTR directs UPF1-dependent mRNA decay in mammalian cells. *Genome Res*, **27**, 407-418.

Chapter 5: A Bioinformatics Approach for the Identification of Telomere-Binding Proteins

5.1. Introduction

Human chromosomes terminate in 10-15 kb stretches of the repetitive telomeric DNA sequence, mainly TTAGGG (1,2). Alongside associated proteins, this DNA forms complexes crucial for shielding chromosome ends from inappropriate DNA repair (3). Telomeres are well-known for their correlation with the aging process. The gradual shortening of telomere length occurs during the division of somatic cells (4). Once it shortens critically (around 3-5 kb in humans), it can initiate a DNA damage response, leading to cell senescence or apoptosis (5). This process restricts unchecked cell growth, offering a formidable defense against tumors (4,6). Yet, many tumor cells bypass this by elevating telomerase activity. Intriguingly, 10-15% of human cancers display minimal telomerase and sustain their telomere length via a mechanism known as alternative lengthening of telomeres (ALT) (7,8).

The telomeric DNA G4 was the first identified biologically relevant G4 structure using single-stranded oligonucleotides that represent the telomere sequences of ciliated protozoa (9). The G4-specific antibody, BG4, has facilitated the *in vivo* visualization of G4 structures, revealing the formation of telomere G4 when combined with fluorescence in situ hybridization targeting telomeric DNA (10). The presence of G-rich sequences in telomeres suggests an evolutionary advantage in the formation of G4 structures within these regions, potentially playing a crucial role in telomere biology (11). It has been hypothesized that the formation of G4 in the single-stranded overhangs of telomeric DNA serves as a protective cap against nucleases (12).

Additionally, several helicases, known to unwind G-quadruplexes, have been demonstrated to be vital for telomere maintenance (13,14).

The stability and functions of telomere are mediated by telomere-associated proteins, which are endowed with specificity for telomeres through their preferential recognition of telomeric DNA sequences. Shelterin, a six-subunit protein complex comprised of TRF1, TRF2, POT1, PAR1, TIN2 and TPP1 in mammalian cells, is a group of most well-studied telomere-associated proteins (3). In shelterin-DNA complex, TRF1 and TRF2 bind directly to double-stranded (ds) portion of telomeric DNA, whereas POT1 binds to single-stranded telomeric overhang.

Aside from the shelterin complex, an increasing number of cellular proteins were identified by various proteomic methods and found to modulate telomere biology through telomere protection, telomeric DNA synthesis, and telomere elongation. For example, Dennis et. al employed biotin-labeled telomeric DNA pulldown to demonstrate evolutionary changes of the shelterin complex and discovered eight zinc finger proteins (ZBTB7A, ZBTB10, ZBTB48, ZNF276, ZNF524, ZNF827, VEZF1 and KLF12) as putative telomere binders (15). However, these methods relying on mass spectrometry may not capture low-abundance telomere-binding proteins. Therefore, alternative methods are needed for effective identification of telomere-binding proteins, which may offer new insights into telomere biology.

In this study, we utilized a bioinformatic method to identify novel telomere-binding proteins. Our approach was based on a comprehensive enrichment analysis of ~250 publicly available ChIP-seq datasets, covering over 130 target proteins. We particularly focus on calculating the enrichment score of zinc finger family proteins, regarding to various telomeric types and repeat lengths. In addition, we examined the telomeric repeat enrichments for proteins

that exhibit a high degree of co-localization with native G4 sites. From this analysis, we identified ZBTB33 as a potential G4-dependent telomere-binding protein that could be further investigated.

5.2. Materials and Methods

Data acquisition and visualization

The ChIP-seq data for proteins of interests were retrieved from the ENCODE portal under TFs ChIP-seq section (16,17). Extremely low read depth, un-replicated and drug treatment experiments were excluded. Unfiltered alignment files from a total of 255 datasets, along with their corresponding control datasets, were downloaded and processed. Samtools were employed to calculate the number of reads containing telomeric sequences, including $(TTAGGG)_n$, $(TAAGGG)_n$, $(TGAGGG)_n$ and $(TCAGGG)_n$ (18). The IP signal was defined as reads per million mapped reads, and the enrichment score was determined as IP signal in experiment versus in control datasets (Figure 5.1). Bigwig files for visualization telomeric regions were generated from unfiltered alignment files and visualized using IGV (19).

G4 overlapping analysis.

BG4 ChIP-seq data for K562 and HepG2 cells were obtained from Sequence Read Archive (SRA) with the accession number of PRNJ60617 (20). BG4 ChIP-seq data were processed following previously published procedures in GRCh38 assembly. TF ChIP-seq data were retrieved from ENCODE portal and IDR thresholded peaks in GRCh38 assembly were used for overlapping analysis. Overlapping percentages between TFs ChIP-seq and BG4 ChIP-seq were calculated using IntervalStats with GRCh38 genome as domain (21).

Aggregation plot

The enrichment profile of target protein's ChIP signal with respect to BG4 peaks was obtained using Deeptools plotprofile function (22). ChIP-seq signals were plotted within the +/- 2000 bps with respect to target protein peak center or BG4 peak centers.

5.3. Results

Zinc finger proteins constitute a family of proteins, each containing one or more zinc finger domains. These domains enable the proteins to bind nucleic acids. Several zinc finger proteins, including ZNF827, have been identified as telomere-binding proteins (23). Additionally, biotin pulldown proteomics experiments have identified several ZNF family proteins. Given these findings, it becomes pertinent to investigate, using a ChIP-seq approach, whether these proteins have telomeric-binding affinities. A total of 34 ZNF proteins were analyzed for enrichment scores associated with various telomeric repeats except for $(TTGGGG)_n$, which cannot be reliably calculated due to extremely low abundance (Figure). Of the proteins analyzed, 11 showed significant enrichment for $(TTAGGG)_4$ repeat sequences, while ZNF124 exhibited a notable depletion of this sequence. Some of these targets also displayed significantly enrichment of telomeric DNA variants (Figure 5.2, Table 5.1). Among these targets, ZNF316 and ZNF24 have overall highest enrichment folds.

ZNF24, also known as ZNF191 or ZSCAN3, is recognized for its indispensable role in regulating key processes during organ development (24). The N-terminus of ZNF24 contains a SCAN domain (named after SRE-ZBP, CTfin51, AW-1 and Number 18 cDNA), facilitating protein-protein interactions, while its C-terminus carries four C2H2-type zinc finger domains (Figure 5.3). Motif analysis have revealed its preference for binding 5'-TCAT-3', however, its affinity to telomere sequence remains to be clarified (25). ZNF316, on the other hand, contains 15 zinc finger domains and is highly disordered (Figure 5.3). The function of ZNF316 in gene

regulation and telomere maintenance is much understudied. Our enrichment analysis suggests potential interactions between ZNF316, ZNF24 and telomere DNAs, providing valuable direction for subsequent biological investigation of these two proteins.

Poxvirus and zinc finger (POZ) and Kruppel-type (POK) proteins, also known as ZBTB (zinc finger and BTB domain) proteins, have been shown to act as transcription regulators implicated in vital biological processes such as organ development, differentiation, and cancer progression (26). Previous studies identified three ZBTB family proteins, ZBTB10, ZBTB40 and ZBTB48, as telomeric DNA-binding proteins (15,27-30). Notably, ZBTB48 has been observed to favor longer telomeres and promote telomere trimming. Therefore, we undertook enrichment analysis for ZBTB family proteins using public available ChIP-seq data.

Our data showed ZBTB48 enrichment in the telomeric regions of HEK293 cells with a $\log_2(IP_{telo}/Input_{telo})$ value of 3.3, which is in line with previously published studies. ZBTB40 also exhibited a significant enrichment of the telomeric DNA sequence. Intriguingly, our bioinformatic exploration also revealed significant enrichments of ZBTB11 and ZBTB33, neither of which have been previously identified as telomere-binding proteins. ZBTB33, in particular, showed the most significant enrichment and exhibited the highest median $\log_2(IP_{telo}/Input_{telo})$ value of 3.7 based on 15 independent ChIP-seq datasets from 7 different cell lines (A549, GM12878, HCT116, HepG2, K562, MCF-7 and SK-N-SH) and 2 liver tissues (from a 4-year old female and a 32-year old male (Figure 5.4). To avoid arbitrary choice of repeat length, we calculated enrichment score in different repeat numbers ranging from 2 to 6 (as analyzed sequencing dataset was acquired in single-end 36 bp reads). We observed a gradual increasing ratio of IP_{telo} versus $Input_{telo}$, which further substantiated ZBTB33's enrichment of telomeric sequences rather than short repeats (Figure 5.5). Furthermore, IGV plots of ZBTB33 displayed high ChIP signal at telomere regions

and co-localized with known-telomere binding proteins TERF1, which substantiates our conclusion of ZBTB33 as a potential telomere-binding protein (Figure 5.6).

ZBTB33, also known as kaiso, has been shown to binds three different motifs including 5'-TCCTGCNA-3', 5'-TCTCGCGAGA-3' and methylated 5'-CGCG-3' (31). Despite ongoing debates over its consensus binding sites, ZBTB33 has been shown to actively participate in transcriptional regulation of tumour-associated genes. Previous studies demonstrated that most of ZBTB33 ChIP-seq peaks overlapped with RNA polymerase II peaks and enriched for active histone modifications (32). Given its pronounced enrichment in the telomeric sequence capable of G4 formation, we aim to delve deeper into G4's potential role in telomere DNA enrichment of ZBTB33.

Our overlapping analysis of ZBTB33 with BG4 ChIP-seq revealed a 44.2% overlapping percentage, with a significance level of $p < 0.01$ (Figure 5.7). Representative IGV plots demonstrated the co-localization of ZBTB33 with endogenous G4 structures (Figure 5.8). Furthermore, profiling of ZBTB33 ChIP signals relative to native G4 sites highlighted a pronounced enrichment, and vice versa. This overlapping analysis suggests that ZBTB33 might be a potential G4-interacting proteins. Combining with previous telomeric enrichment analysis, we postulate ZBTB33 to be a G4-dependent telomeric DNA-interacting protein.

5.4. Conclusions

Telomeres are specialized structures located at the ends of chromosomes, protecting them from degradation, fusion and inappropriate recombination (33). Because the essential role that telomeres play in maintaining chromosome integrity, cellular aging and cancer, its interacting proteins has drawn considerable attention for their role in maintaining the structure and function of telomeres (34).

Most methods in identifying telomere-interacting proteins rely on proteomics analysis via mass spectrometry. A common strategy includes utilizing biotin-labeled telomeric repeat oligonucleotides for nuclear pulldowns (15). Through this methodology, researchers have identified several zinc finger proteins as telomere-binding entities and elucidated some of their roles in telomere maintenance.

In this study, we developed a novel method to identify candidate telomere-interacting proteins through bioinformatics analysis of publicly available ChIP-seq date. We extensively analyzed ChIP-seq datasets of zinc finger proteins, in particular ZBTB family proteins, assessing their immunoprecipitation enrichment of telomeric repeat sequences. Among these targets, ZNF316, ZNF24 and ZBTB33 exhibited significant enrichment with top-ranking scores. An in-depth evaluation of ZBTB33 genome-wide distribution revealed its co-localization with native G4 structures.

Numerous studies have delved into the role of ZBTB33 in transcription regulation, especially in tumour-related processes (31). ZBTB33 has been shown to activate tumor cell invasion and metastasis through TGFβ signaling, miR-200 family and E-cadherin expression (35-37). Comprehensive analysis of TCGA and GTEx database revealed an elevated expression of ZBTB33 across nearly all tumor types (38). Significant up-regulations were observed in 9 tumour types when compared with normal tissues (Figure 5.8). Given its importance in cancer progression, our enrichment analysis, which pinpoints ZBTB33 as putative telomere binding proteins, may shed light on novel mechanism of this protein in cancer biology.

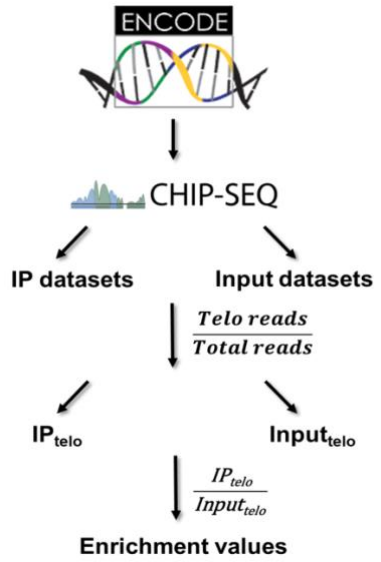

Figure 5.1. A schematic diagram showing the bioinformatic workflow for uncovering telomere binding proteins using publicly available ChIP-seq data.

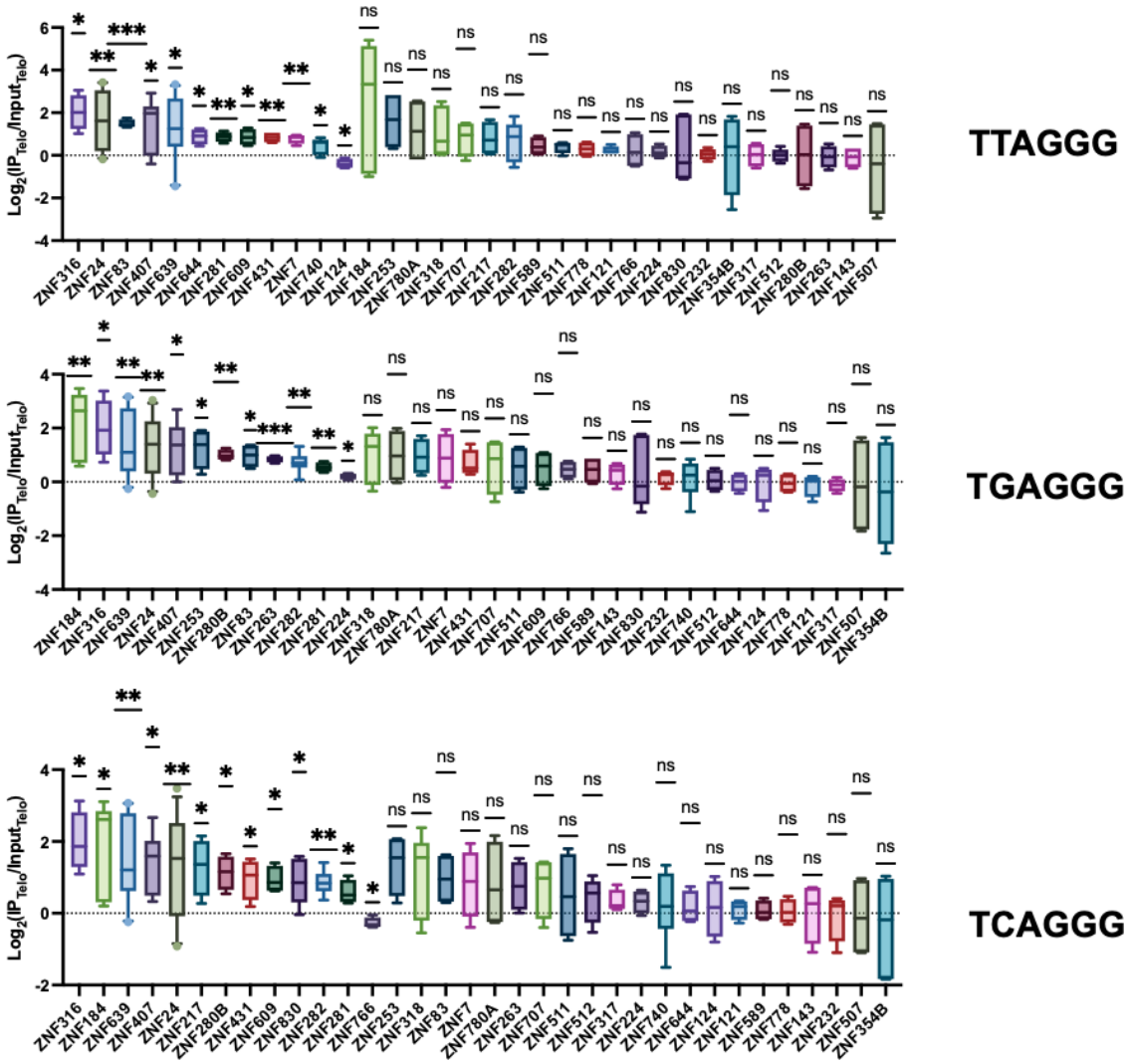

Figure 5.2. Enrichment values of ZNF proteins at $(TTAGGG)_4$, $(TGAGGG)_4$ and $(TCAGGG)_4$ with at least three independent experiments. *p* values were calculated using the one-sample t and Wilcoxon test. *, $0.01 < p < 0.05$; **, $0.001 < p < 0.01$; ***, $0.0001 < p < 0.001$; ns, not significant.

	TTAGGG	TGAGGG	TCAGGG
ZNF124	✓		
ZNF184		✓	✓
ZNF217			✓
ZNF224		✓	
ZNF24	✓	✓	✓
ZNF253		✓	
ZNF263		✓	
ZNF280B		✓	✓
ZNF281	✓	✓	✓
ZNF282		✓	✓
ZNF316	✓	✓	✓
ZNF40	✓		
ZNF407	✓	✓	✓
ZNF431	✓		✓
ZNF609	✓		✓
ZNF639	✓	✓	✓
ZNF644	✓		
ZNF7	✓		
ZNF766			✓
ZNF83	✓	✓	
ZNF830			✓

Table 5.1. Overview of significant targets across three telomeric repeat variants. Proteins depleted in corresponding sequences are highlighted in red, while those enriched in all three variants are shaded in green.

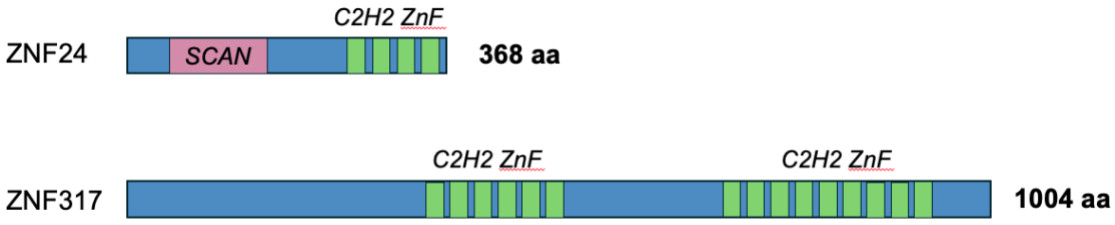

Figure 5.3. Schematic diagrams depicting the domain structures of ZNF24 and ZNF317.

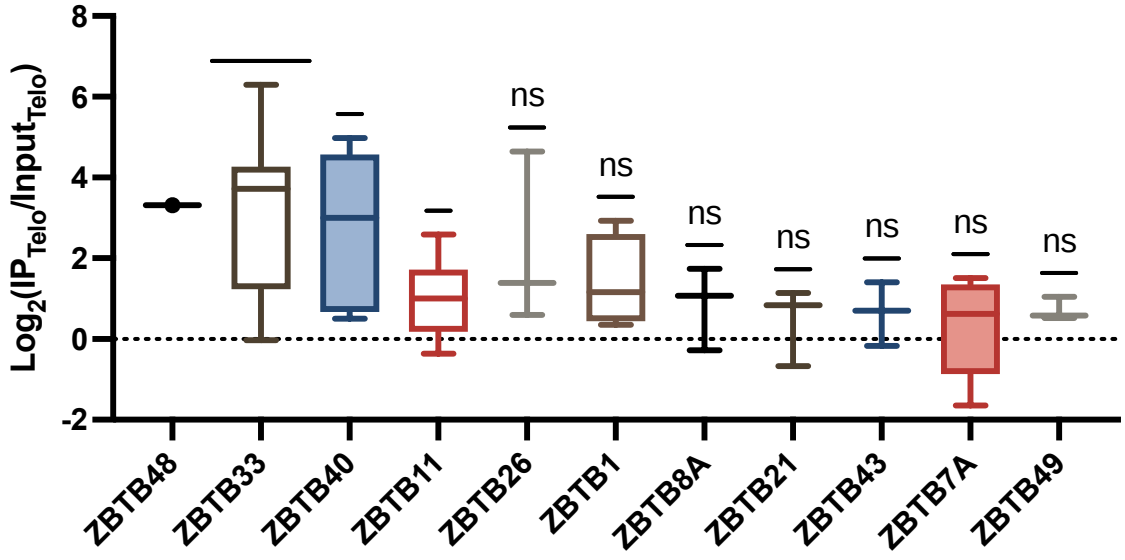

Figure 5.4. Enrichment values of ZBTB proteins at $(TTAGGG)_4$ repeats with at least three independent experiments. *p* values were calculated using the one-sample *t* and Wilcoxon test. *, $0.01 < p < 0.05$; ****, $p < 0.0001$; ns, not significant.

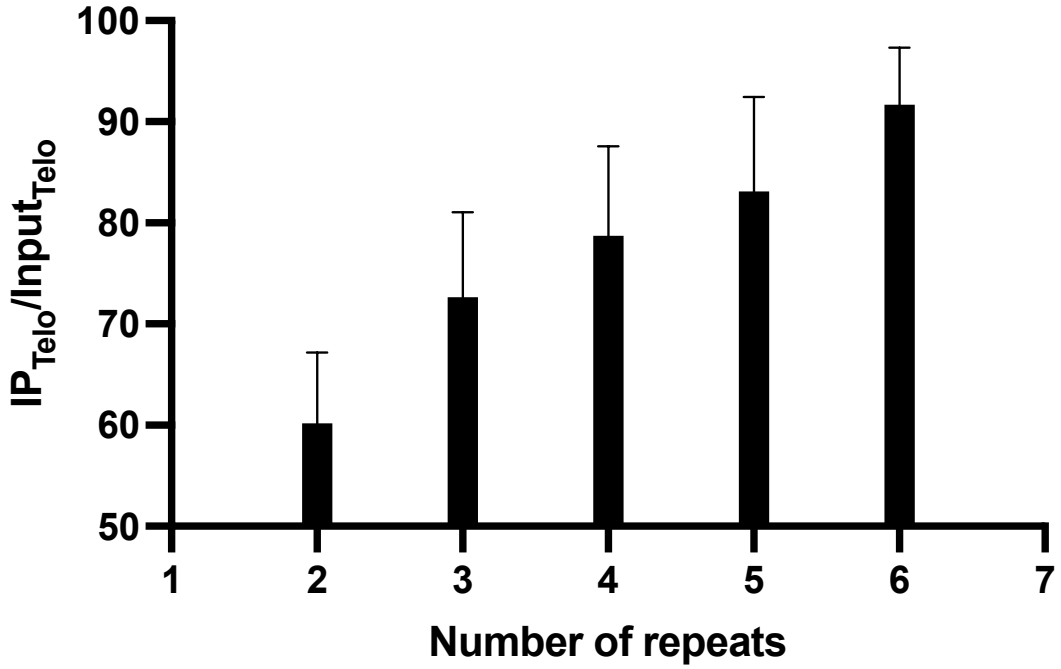

Figure 5.5. Calculation of the enrichment score for ZBTB33 (ENCSR000BNA) across repeat numbers ranging from 2 to 6. A progressive increase with the number of repeats suggests ZBTB33's preference for longer telomeric DNA over short TTAGGG sequences.

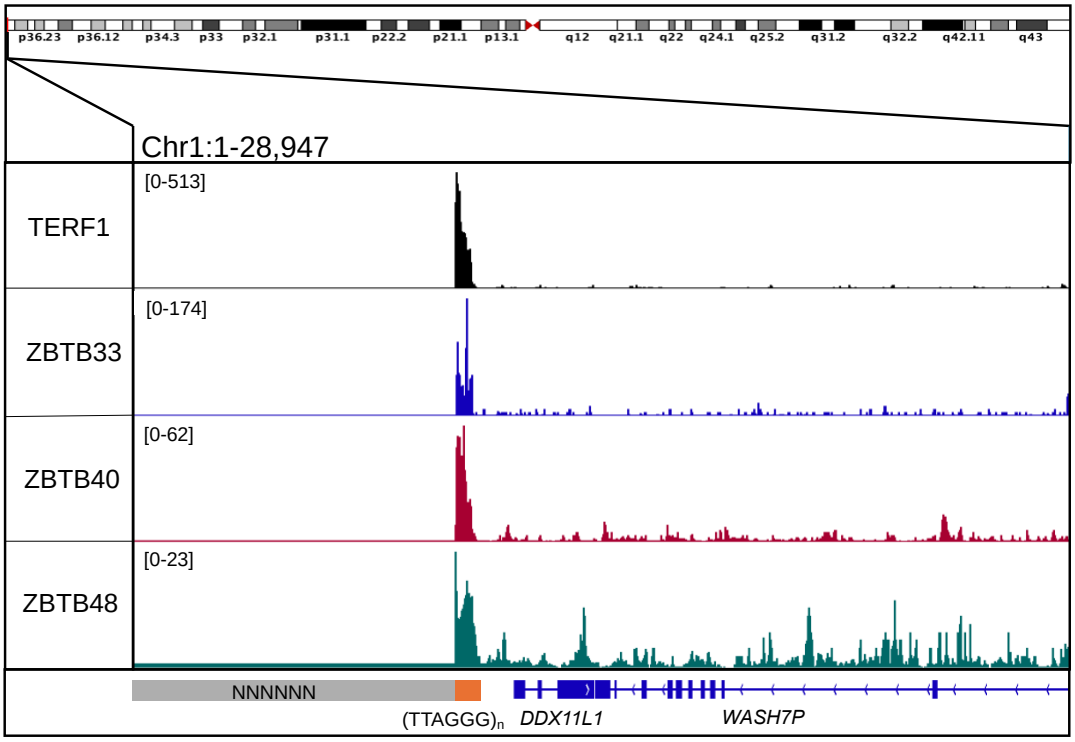

Figure 5.6. IGV plot showing the enrichment of ZBTB proteins at telomere region. TERF1, a known telomere binding protein, serves as a reference.

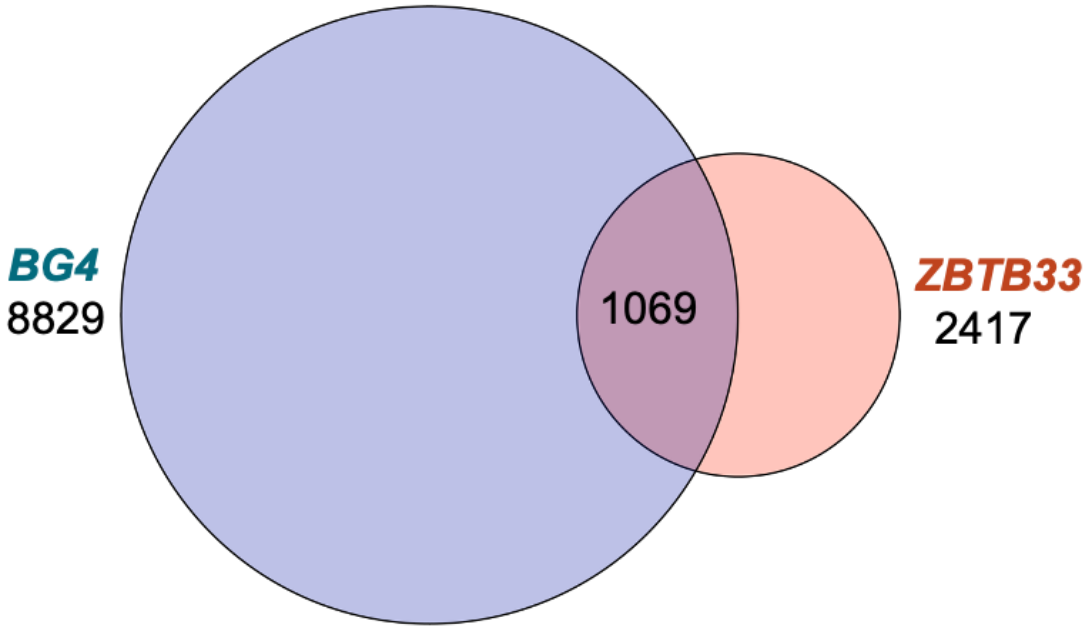

Figure 5.7. A Venn diagram showing overlapping between ZBTB33 and BG4 ChIP-seq.

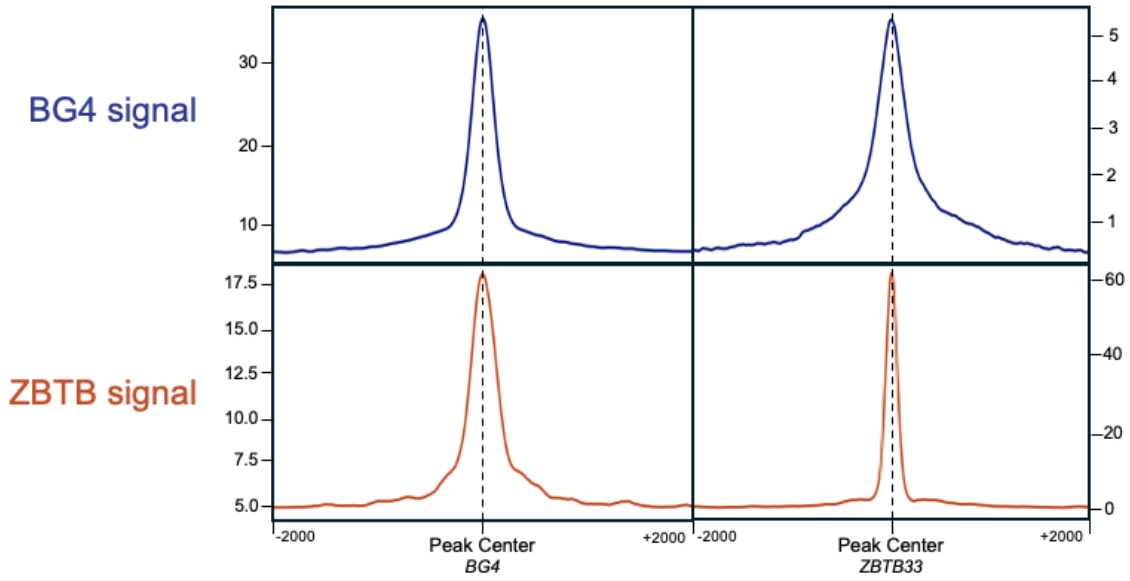

Figure 5.8. Representative enrichment profiles of ZBTB33 and BG4 with respect to corresponding peak centers.

Figure 5.9. Expression profiles of ZBTB33 in all tumor samples compared to paired normal tissues. Each dot represents the expression level of an individual sample. Tumor types showing significant differences are labeled in red.

5.5. References

1. de Lange, T., Shiue, L., Myers, R.M., Cox, D.R., Naylor, S.L., Killery, A.M. and Varmus, H.E. (1990) Structure and variability of human chromosome ends. *Mol Cell Biol*, **10**, 518-527.

2. Moyzis, R.K., Buckingham, J.M., Cram, L.S., Dani, M., Deaven, L.L., Jones, M.D., Meyne, J., Ratliff, R.L. and Wu, J.R. (1988) A highly conserved repetitive DNA sequence, (TTAGGG)n, present at the telomeres of human chromosomes. *Proc. Natl. Acad. Sci. U. S. A.*, **85**, 6622-6626.

3. De Lange, T. (2005) Shelterin: the protein complex that shapes and safeguards human telomeres. *Genes Dev.*, **19**, 2100-2110.

4. Shay, J.W. (2016) Role of telomeres and telomerase in aging and cancer. *Cancer Discov.*, **6**, 584-593.

5. Harley, C.B., Futcher, A.B. and Greider, C.W. (1990) Telomeres shorten during ageing of human fibroblasts. *Nature*, **345**, 458-460.

6. Maciejowski, J. and de Lange, T. (2017) Telomeres in cancer: tumour suppression and genome instability. *Nature reviews Molecular cell biology*, **18**, 175-186.

7. Cong, Y.-S., Wright, W.E. and Shay, J.W. (2002) Human telomerase and its regulation. *Microbiology and molecular biology reviews*, **66**, 407-425.

8. Cesare, A.J. and Reddel, R.R. (2010) Alternative lengthening of telomeres: models, mechanisms and implications. *Nature reviews genetics*, **11**, 319-330.

9. Henderson, E., Hardin, C.C., Walk, S.K., Tinoco, I., Jr. and Blackburn, E.H. (1987) Telomeric DNA oligonucleotides form novel intramolecular structures containing guanine-guanine base pairs. *Cell*, **51**, 899-908.

10. Biffi, G., Tannahill, D., McCafferty, J. and Balasubramanian, S. (2013) Quantitative visualization of DNA G-quadruplex structures in human cells. *Nat. Chem.*, **5**, 182-186.

11. Tran, P.L.T., Mergny, J.-L. and Alberti, P. (2011) Stability of telomeric G-quadruplexes. *Nucleic Acids Res.*, **39**, 3282-3294.

12. Smith, J.S., Chen, Q., Yatsunyk, L.A., Nicoludis, J.M., Garcia, M.S., Kranaster, R., Balasubramanian, S., Monchaud, D., Teulade-Fichou, M.-P. and Abramowitz, L. (2011) Rudimentary G-quadruplex–based telomere capping in Saccharomyces cerevisiae. *Nat. Struct. Mol. Biol.*, **18**, 478-485.

13. Drosopoulos, W.C., Kosiyatrakul, S.T. and Schildkraut, C.L. (2015) BLM helicase facilitates telomere replication during leading strand synthesis of telomeres. *J. Cell Biol.*, **210**, 191-208.

14. Wu, W., Bhowmick, R., Vogel, I., Özer, Ö., Ghisays, F., Thakur, R.S., Sanchez de Leon, E., Richter, P.H., Ren, L. and Petrini, J.H. (2020) RTEL1 suppresses G-quadruplex-associated R-loops at difficult-to-replicate loci in the human genome. *Nat. Struct. Mol. Biol.*, **27**, 424-437.

15. Kappei, D., Scheibe, M., Paszkowski-Rogacz, M., Bluhm, A., Gossmann, T.I., Dietz, S., Dejung, M., Herlyn, H., Buchholz, F., Mann, M. *et al.* (2017) Phylointeractomics reconstructs functional evolution of protein binding. *Nat Commun*, **8**, 14334.

16. Consortium, E.P. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57-74.

17. Davis, C.A., Hitz, B.C., Sloan, C.A., Chan, E.T., Davidson, J.M., Gabdank, I., Hilton, J.A., Jain, K., Baymuradov, U.K., Narayanan, A.K. *et al.* (2018) The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res.*, **46**, D794-D801.

18. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. and Genome Project Data Processing, S. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078-2079.

19. Thorvaldsdóttir, H., Robinson, J.T. and Mesirov, J.P. (2012) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, **14**, 178-192.

20. Spiegel, J., Cuesta, S.M., Adhikari, S., Hansel-Hertsch, R., Tannahill, D. and Balasubramanian, S. (2021) G-quadruplexes are transcription factor binding hubs in human chromatin. *Genome Biol.*, **22**, 117.

21. Chikina, M.D. and Troyanskaya, O.G. (2012) An effective statistical evaluation of ChIPseq dataset similarity. *Bioinformatics*, **28**, 607-613.

22. Ramirez, F., Ryan, D.P., Gruning, B., Bhardwaj, V., Kilpert, F., Richter, A.S., Heyne, S., Dundar, F. and Manke, T. (2016) deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.*, **44**, W160-165.

23. Conomos, D., Reddel, R.R. and Pickett, H.A. (2014) NuRD–ZNF827 recruitment to telomeres creates a molecular scaffold for homologous recombination. *Nat. Struct. Mol. Biol.*, **21**, 760-770.

24. Khalfallah, O., Faucon-Biguet, N., Nardelli, J., Meloni, R. and Mallet, J. (2008) Expression of the transcription factor Zfp191 during embryonic development in the mouse. *Gene Expression Patterns*, **8**, 148-154.

25. Sandelin, A., Alkema, W., Engstrom, P., Wasserman, W.W. and Lenhard, B. (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.*, **32**, D91-94.

26. Lee, S.U. and Maeda, T. (2012) POK/ZBTB proteins: an emerging family of proteins that regulate lymphoid development and function. *Immunol. Rev.*, **247**, 107-119.

27. Bluhm, A., Viceconte, N., Li, F., Rane, G., Ritz, S., Wang, S., Levin, M., Shi, Y., Kappei, D. and Butter, F. (2019) ZBTB10 binds the telomeric variant repeat TTGGGG and interacts with TRF2. *Nucleic Acids Res.*, **47**, 1896-1907.

28. Wang, S., Xu, Z., Li, M., Lv, M., Shen, S., Shi, Y. and Li, F. (2023) Structural insights into the recognition of telomeric variant repeat TTGGGG by Kruppel-like C2H2 finger protein ZBTB10. *J. Biol. Chem.*, 102918.

29. Jahn, A., Rane, G., Paszkowski-Rogacz, M., Sayols, S., Bluhm, A., Han, C.T., Draskovic, I., Londono-Vallejo, J.A., Kumar, A.P., Buchholz, F. *et al.* (2017) ZBTB48 is both a vertebrate telomere-binding protein and a transcriptional activator. *EMBO Rep*, **18**, 929-946.

30. Li, J.S., Miralles Fuste, J., Simavorian, T., Bartocci, C., Tsai, J., Karlseder, J. and Lazzerini Denchi, E. (2017) TZAP: A telomere-associated protein involved in telomere length control. *Science*, **355**, 638-641.

31. Pierre, C.C., Hercules, S.M., Yates, C. and Daniel, J.M. (2019) Dancing from bottoms up–roles of the POZ-ZF transcription factor Kaiso in cancer. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer*, **1871**, 64-74.

32. Blattler, A., Yao, L., Wang, Y., Ye, Z., Jin, V.X. and Farnham, P.J. (2013) ZBTB33 binds unmethylated regions of the genome associated with actively expressed genes. *Epigenetics & chromatin*, **6**, 1-18.

33. Shay, J.W. and Wright, W.E. (2019) Telomeres and telomerase: three decades of progress. *Nature Reviews Genetics*, **20**, 299-309.

34. Martínez, P. and Blasco, M.A. (2011) Telomeric and extra-telomeric roles for telomerase and the telomere-binding proteins. *Nature Reviews Cancer*, **11**, 161-176.

35. Jones, J., Wang, H., Karanam, B., Theodore, S., Dean-Colomb, W., Welch, D.R., Grizzle, W. and Yates, C. (2014) Nuclear localization of Kaiso promotes the poorly differentiated phenotype and EMT in infiltrating ductal carcinomas. *Clin. Exp. Metastasis*, **31**, 497-510.

36. Abisoye-Ogunniyan, A., Lin, H., Ghebremedhin, A., Salam, A.B., Karanam, B., Theodore, S., Jones-Trich, J., Davis, M., Grizzle, W. and Wang, H. (2018) Transcriptional repressor Kaiso promotes epithelial to mesenchymal transition and metastasis in prostate cancer through direct regulation of miR-200c. *Cancer Lett.*, **431**, 1-10.

37. Bassey-Archibong, B., Kwiecien, J., Milosavljevic, S., Hallett, R., Rayner, L., Erb, M., Crawford-Brown, C., Stephenson, K., Bédard, P. and Hassell, J. (2016) Kaiso depletion attenuates transforming growth factor-β signaling and metastatic activity of triple-negative breast cancer cells. *Oncogenesis*, **5**, e208-e208.

38. Tang, Z., Kang, B., Li, C., Chen, T. and Zhang, Z. (2019) GEPIA2: an enhanced web server for large-scale expression profiling and interactive analysis. *Nucleic Acids Res.*, **47**, W556-W560.

Chapter 6: Concluding Remarks and Future Direction

In this dissertation, we rigorously examined the functions of DNA G4 in the epigenetic modulation of transcription. Our investigations uncovered a G4 overlap with transcription factors and yielded an interaction network of putative G4-interacting proteins. We further delved into the relationship between G4 structures and histone marks. Through a combined analysis of histone modification enzymes, we proposed a G4-dependent modulation of histone modifications.

We also performed a comprehensive overlapping analysis incorporating previously published RNAPII ChIA-PET and BG4 ChIP-seq datasets. Notably, we detected a robust positive association between RNAPII-associated DNA loops and chromatin G4 structures. Through HiChIP-seq and RNA-seq with small-molecule G4 ligand treatment, we underscore the pivotal role of DNA G4 in RNAPII-mediated DNA looping and transcriptional modulation.

To delineate the functional roles of RNA G4s, we adopted a bioinformatic strategy, analyzing overlap between rG4-seq peaks and those discerned in >230 eCLIP-seq datasets for RNA-binding proteins from the ENCODE project. This led to the identification of numerous prospective rG4-binding proteins. Among these targets, we validated G3BP1 as a *bona fide* RNA G4-binding proteins. We further examined its role in translation regulation and RNA metabolism.

Lastly, we introduced an innovative methodology to discern putative telomere-binding proteins via ChIP-seq data. Key targets, namely ZNF24, ZNF316, and ZBTB33, manifested pronounced enrichment at telomeric repeat sequences. Detailed scrutiny of ZBTB33 pinpointed potential G4-dependent telomere-binding dynamics, broadening our understanding of proteins integral to telomere biology.

Throughout this dissertation, bioinformatic analysis has been a pivotal foundation for raising biological questions and a crucial tool for deciphering intricate mechanisms. While DNA and RNA G4s have been postulated to regulate essential biological processes (1-3), the absence of genome-wide localization for these secondary structures impedes the elucidation of detail molecular mechanisms behind G4 functions. Our studies comprehensively explored the correlation between native G4 localizations and important regulatory elements, including transcription factor binding sites, DNA looping anchors, RNA-binding protein binding sites, and telomeric repeat sequences. Leveraging the comprehensive ENCODE data repository, we unveiled novel roles of G4 in transcription regulation, identified new RNA G4-binding proteins, and highlighted potential telomere binding proteins.

With the development of G4 biology and the advancement of sequencing methods, a plethora of unexplored questions can be addressed via next-generation sequencing coupled with bioinformatics analysis. It would be interesting to explore how G4 structures, as regulatory elements, are modulated by environmental exposure. Along this line, guanine oxidation happens from the action of reactive oxygen species induced either by regular cellular metabolism or external factors like radiation (4,5). The primary

outcome, 8-oxo-7,8-dihydroguanine (8-oxoG), can lead to mutations if not appropriately repaired (6-8). Due to their guanine-rich nature, DNA G4 structures are vulnerable to oxidation (9). Intriguingly, several studies indicated that guanine oxidation in G4-forming sequences can stimulate DNA repair and activate downstream gene expression. For instance, oxidation in promoters of VEGF, NTHL1, and PCNA, which all contain G4-forming sequences, increased downstream luciferase expression in reporter genes (10,11). An alternative G4 formation model, suggesting that G4 obstructs the typical nuclease function of APE1 and facilitates transcription factor recruitment, was proposed (12). Yet, these investigations were conducted using plasmid systems with inserted G4-forming sequences. A genome-wide evaluation of G4's role during oxidative stress can be achieved by merging BG4 ChIP-seq with the single-nucleotide resolution mapping of 8-oxoG (13). Analyzing these datasets concurrently may shed light on novel mechanisms through which DNA G4 structures modulate important cellular processes.

Beyond identifying potential RNA G4-binding proteins, genome-wide mapping of RNA G4 structures can also assist in understanding the interplay between these secondary structures and epitranscriptomics. Chemical modifications within transcripts, termed epitranscriptomics, have been shown to influence numerous aspects of RNA metabolism and function, from reshaping RNA structures and dictating RNA-protein interactions to determining RNA stability and controlling translation (14,15). Numerous studies have linked RNA modifications with RNA G4 structures. G4RP-MS investigations indicated that RNA G4 forming sequences could attract 'writers', 'readers', and 'erasers' of RNA modifications, including FXR1, ADAR, and BUD23 (16,17). Furthermore, bioinformatic

analyses aligning RNA G4 with m^6A suggest a cooperative regulatory role between RNA modifications and RNA G4s in processes such as pre-mRNA alternative splicing or viral post-transcriptional gene expression (18,19). With the advancement of RNA modification mapping methodologies, integrating bioinformatics analysis with native RNA G4 localization can clarify the cooperative roles of the epitranscriptome and RNA G4 structures.

Our examination of potential telomere-binding proteins could be expanded to include the identification of TERRA-interacting proteins. TERRA, a transcribed form of telomere DNA, has demonstrated its role in modulating chromatin states within telomere regions and in maintaining telomere length (20). Preliminary bioinformatic analysis of RNA-binding protein eCLIP-seq datasets revealed minimal presence of the UUAGGG sequence, even in datasets for known TERRA-associated proteins. This may be attributed to secondary structure interference during cDNA synthesis or the length distribution of TERRA. We look forward to advancements in RNA-protein interaction profiling, which will not only assist in identifying new TERRA-interacting proteins, but also shed light on previously unknown G4-dependent functions of RNA-binding proteins.

In conclusion, this dissertation introduces a novel perspective on the role of G4 in various biological pathways, encompassing transcription, translation, and telomere maintenance. Our research not only highlights G4 structures as multi-faceted regulatory elements within the complex cellular milieu, but also connects them with other essential regulatory mechanisms to establish an interconnected regulatory network. Further investigation delving deeper into the relationship between G4 structure and cellular

processes, such as the DNA damage response and epitranscriptomics regulation, could enhance our comprehension of the diverse role of G4, which paves the way for groundbreaking insights in therapeutic research.

6.1. References

1. Robinson, J., Raguseo, F., Nuccio, S.P., Liano, D. and Di Antonio, M. (2021) DNA G-quadruplex structures: more than simple roadblocks to transcription? *Nucleic Acids Res.*, **49**, 8419-8431.

2. Spiegel, J., Adhikari, S. and Balasubramanian, S. (2020) The Structure and Function of DNA G-Quadruplexes. *Trends Chem*, **2**, 123-136.

3. Dumas, L., Herviou, P., Dassi, E., Cammas, A. and Millevoi, S. (2021) G-Quadruplexes in RNA Biology: Recent Advances and Future Directions. *Trends Biochem Sci*, **46**, 270-283.

4. Ward, J.F. (2000) Complexity of damage produced by ionizing radiation. *Cold Spring Harb. Symp. Quant. Biol.*, **65**, 377-382.

5. Markkanen, E. (2017) Not breathing is not an option: How to deal with oxidative DNA damage. *DNA Repair (Amst)*, **59**, 82-105.

6. Steenken, S. and Jovanovic, S.V. (1997) How Easily Oxidizable Is DNA? One-Electron Reduction Potentials of Adenosine and Guanosine Radicals in Aqueous Solution. *Journal of the American Chemical Society*, **119**, 617-618.

7. Cadet, J., Wagner, J.R., Shafirovich, V. and Geacintov, N.E. (2014) One-electron oxidation reactions of purine and pyrimidine bases in cellular DNA. *Int J Radiat Biol*, **90**, 423-432.

8. Grollman, A.P. and Moriya, M. (1993) Mutagenesis by 8-oxoguanine: an enemy within. *Trends Genet.*, **9**, 246-249.

9. Clark, D.W., Phang, T., Edwards, M.G., Geraci, M.W. and Gillespie, M.N. (2012) Promoter G-quadruplex sequences are targets for base oxidation and strand cleavage during hypoxia-induced transcription. *Free Radic. Biol. Med.*, **53**, 51-59.

10. Fleming, A.M., Ding, Y. and Burrows, C.J. (2017) Oxidative DNA damage is epigenetic by regulating gene transcription via base excision repair. *Proc. Natl. Acad. Sci. U. S. A.*, **114**, 2604-2609.

11. Broxson, C., Hayner, J.N., Beckett, J., Bloom, L.B. and Tornaletti, S. (2014) Human AP endonuclease inefficiently removes abasic sites within G4 structures compared to duplex DNA. *Nucleic Acids Res.*, **42**, 7708-7719.

12. Redstone, S.C.J., Fleming, A.M. and Burrows, C.J. (2019) Oxidative Modification of the Potential G-Quadruplex Sequence in the PCNA Gene Promoter Can Turn on Transcription. *Chem. Res. Toxicol.*, **32**, 437-446.

13. Wu, J., McKeague, M. and Sturla, S.J. (2018) Nucleotide-Resolution Genome-Wide Mapping of Oxidative DNA Damage by Click-Code-Seq. *J. Am. Chem. Soc.*, **140**, 9783-9787.

14. Wiener, D. and Schwartz, S. (2021) The epitranscriptome beyond m(6)A. *Nat. Rev. Genet.*, **22**, 119-131.

15. Delaunay, S. and Frye, M. (2019) RNA modifications regulating cell fate in cancer. *Nat. Cell Biol.*, **21**, 552-559.

16. Edupuganti, R.R., Geiger, S., Lindeboom, R.G.H., Shi, H., Hsu, P.J., Lu, Z., Wang, S.Y., Baltissen, M.P.A., Jansen, P., Rossa, M. *et al.* (2017) N(6)-methyladenosine (m(6)A) recruits and repels proteins to regulate mRNA homeostasis. *Nat. Struct. Mol. Biol.*, **24**, 870-878.

17. Arguello, A.E., DeLiberto, A.N. and Kleiner, R.E. (2017) RNA Chemical Proteomics Reveals the N(6)-Methyladenosine (m(6)A)-Regulated Protein-RNA Interactome. *J. Am. Chem. Soc.*, **139**, 17249-17252.

18. Jara-Espejo, M., Fleming, A.M. and Burrows, C.J. (2020) Potential G-Quadruplex Forming Sequences and N(6)-Methyladenosine Colocalize at Human Pre-mRNA Intron Splice Sites. *ACS Chem. Biol.*, **15**, 1292-1300.

19. Fleming, A.M., Nguyen, N.L.B. and Burrows, C.J. (2019) Colocalization of m(6)A and G-Quadruplex-Forming Sequences in Viral RNA (HIV, Zika, Hepatitis B, and SV40) Suggests Topological Control of Adenosine N (6)-Methylation. *ACS Cent Sci*, **5**, 218-228.

20. Bettin, N., Oss Pegorar, C. and Cusanelli, E. (2019) The Emerging Roles of TERRA in Telomere Maintenance and Genome Stability. *Cells*, **8**.