UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI BIOLOGIA

Corso di Laurea magistrale in Molecular Biology

TESI DI LAUREA

STUDY OF AN ARTIFICIAL G4-BINDING PROTEIN AS A TOOL TO DEFINE THE PROTEOME ASSEMBLED AROUND G-QUADRUPLEX FORMING SEQUENCES AND EFFICIENT REPLICATION ORIGINS IN DT40 CHICKEN CELLS

Relatore: Prof.ssa Chiara Rampazzo
Dipartimento di Biologia
Correlatore: Dott.ssa Kathrin Marheineke
Team "Chromosomal Domains and Replication"
Jacques Monod Institute

Laureanda: Valentina Artuso

ANNO ACCADEMICO 2024/2025

Table of contents

1 Abstract

Life goes on thanks to the ability of single cells to divide and generate new ones; for this process to happen, cells need to replicate their genome. DNA replication starts only in some specific points of the genome of eukaryotic cells, at origins of replication. Putative G-Quadruplex forming Sequences (PQSs), able to form G-quadruplexes (G4s), a DNA non-canonical secondary structure, are among the *cis*-elements necessary for the process.

G4s can be useful to define the proteome at the level of origins of replication; therefore, the aim of this internship is to fine-tune an approach that relies on an artificial G4-binding protein, G4P, able to recognize structured G4s *in vivo,* to identify proteins involved in origin function in DT40 chicken cells.

The host lab previously generated a clone expressing a fusion protein FLAG-G4P-APEX2, and I performed a series of experiments to validate the induction and the use of G4P. I demonstrated that G4P is expressed upon induction and localized to the nucleus, I could see that genome wide, G4P binds often at CGIs and in correspondence of replication origins; I explored the proteome around G4s showing an enrichment of G4BPs, G4 helicases, proteins associated with replication functions and G4s.

2 Introduction

2.1 DNA replication

DNA replication is a crucial process common to all prokaryotic and eukaryotic cells, which allows to faithfully duplicate the genetic information; although there are some differences between the two, replication starts at replication origins, specific points in the genome characterized by essential *cis*-elements and *trans*-factors able to ensure cell duplication (Prioleau and MacAlpine, 2016). The definition of these elements derives from the replicon model (Jacob et al., 1963), which states that the genome contains replicons, so domains of a dimension ranging between 30 and 450 kb, copied by the bidirectional movement of the replication fork; each replicon is indeed defined and replicated by a *trans*-acting initiator and a *cis*-acting replicator. DNA replication can be divided in three main phases: initiation, elongation and termination, respectively the DNA is unwound to allow the formation of the replication forks, Y-shaped DNA structures where the replication machinery is inserted, the RNA templates are positioned there to start the new filaments synthesis and when the two opposing forks meet the synthesis stops (Nasheuer and Meaney, 2024). Replication origins are defined epigenetically, as GC-rich and associated with promoters, but they do not depend on the DNA sequence; to go more into the specifics of the initiation phase, recent evidence confirms that at each replication origin we find the initiator origin recognition complex (ORC) able to bind DNA at replication forks and recruit Mcm2-7, the replicative helicase, that sequentially binds Cdc45 and GINS therefore forming the active helicase complex (Nasheuer and Meaney, 2024). G4 structures may indeed be important for the unwinding of negatively supercoiled DNA therefore rendering easier the loading of ORC to the open chromatin (Besnard et al., 2012).

The regulation of the initiation process involves more than 50 different proteins and is conserved in eukaryotes. More specifically (Figure 2.1), in early G1 phase, ORC binds origins and recruits Cdc6 and Mcm2-7, that is the Pre-RC, pre-replicative complex, composed of two Mcm2-7 hexamers, Cdc6 and Cdt1; at this moment the origin is potentially ready to fire, therefore, to start replicating, but this is not a direct consequence, as a matter of fact, more pre-RCs are loaded in G1 phase than origins activated during S phase in each cell, allowing cells to adapt their replication program upon replication stress and during differentiation and development in multicellular organisms. Cdc6 is phosphorylated thus degraded and Ctd1 binds geminin to form a stable but inactive complex. At the G1/S phase border the pre-RCs are activated by S phase specific kinases (DDK, Cyclin/Cdk) and other loading factors (TobBP1-Donson and Treslin/TICCR-MTBP) which leads to the establishment of the pre-initiation complex (pre-IC) and to the loading of Cdc45 and GINS onto Mcm2-7, forming the active helicase complex, CMG; this can be reversed with the dephosphorylation by

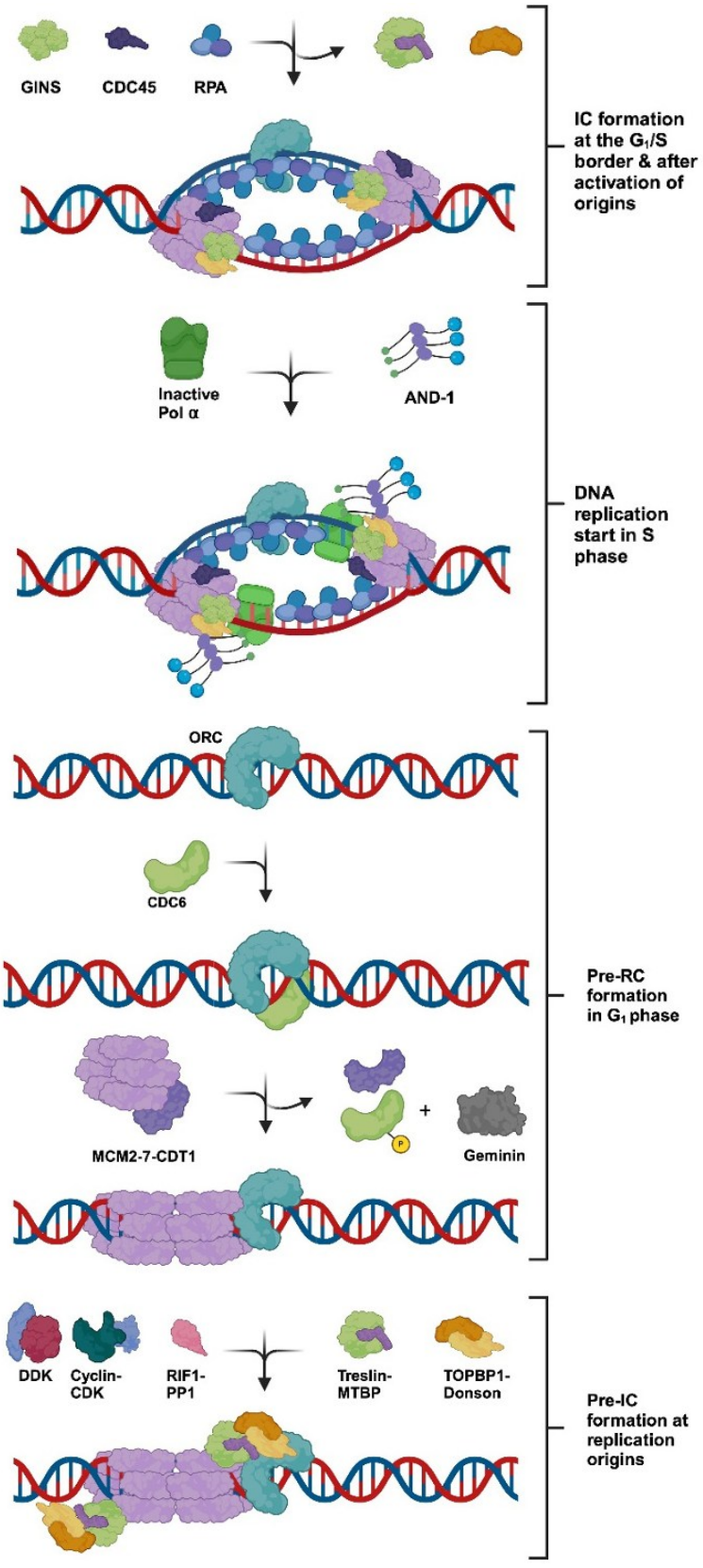
GINS
CDC45
RPA
IC formation at the G1/S border & after activation of origins
Inactive Pol α
AND-1
DNA replication start in S phase
ORC
CDC6
Pre-RC formation in G1 phase
MCM2-7-CDT1
P
Geminin
DDK
Cyclin-CDK
RIF1-PP1
Treslin-MTBP
TOPBP1-Donson
Pre-IC formation at replication origins

Figure 2.1. Overview of the four steps during DNA replication initiation process. 1) Pre-RC formation in G1 phase. 2) Pre-IC formation at replication origins. 3) IC formation at the G1/S border & after activation of origins. 4) DNA replication start in S-phase. (For details, see text**.)** (Nasheuer and Meaney, 2024)

Rif1-PP1, blocking the activation of the complex. The helicase unwinds the DNA and forms the two replication forks stabilizing ssDNA with help of RPA and Cdc45. Ultimately, the primase, DNA polymerase α is loaded to synthesize the RNA primer before DNA polymerase δ and ε continue to synthesize leading and lagging strand, respectively.

Replicons are grouped in bigger constant replication timing domains, conserved among vertebrate species and characterized by different replication timings, therefore they are defined as early, mid or late replicating domains (Figure 2.2). Each of them contains different classes of genes, for example in early regions we find most protein coding genes where there's a vast number of efficient origins; on the contrary, late regions are characterized by a cascade of stochastic events due to the inefficiency of origins. They are separated between each other by an insulator protein, CTCF (CCCTC-binding factor) that determines the topologically associated domains (TADs) boundaries (Prioleau and MacAlpine, 2016).

TADs are regions in which the chromatin is divided into and where the DNA sequences interact only between each other inside each TAD. This is why they must be separated by insulator proteins like CTCF, in order for each of them to be replicated with their own timing. Indeed, for example, late replicating domains are associated to the nuclear lamina (LADs), the inner nuclear membrane, therefore they are closer to the outer part of the nucleus, where we find the majority of the chromatin in the closed state (Prioleau and MacAlpine, 2016). Open chromatin on the other hand, is positioned mostly in the center of the nucleus, where we therefore find early replicating domains, that must be open in order to allow the binding of ORC and thus potentially fire the origin (Valton and Prioleau, 2016). Another kind of TAD that we can find, are those associated to Rif1 in late zones, termed RADs and with a role in dephosphorylating ORC1, therefore stabilizing it for the promotion of initiation.

Replication timing is fundamental to inhibit replication from happening more than once at each cell cycle and to ensure an accurate information transfer, otherwise genomic instability may occur, leading to cancer; indeed, it starts at a specific timepoint during the cell cycle, that is S-phase. The cell cycle is divided into G1, S, G2 and M-phases, where G1 and G2 are needed for the cell itself to make sure to be able to support S and M-phase, while S-phase is when the DNA is replicated and M-phase when the chromosomes, now duplicated, can segregate and the cell divides. With the progression of the cell cycle, each phase change is controlled by the checkpoints, regulated by

cyclin-dependent kinases (CDKs) that phosphorylate target proteins allowing the cell cycle progression.

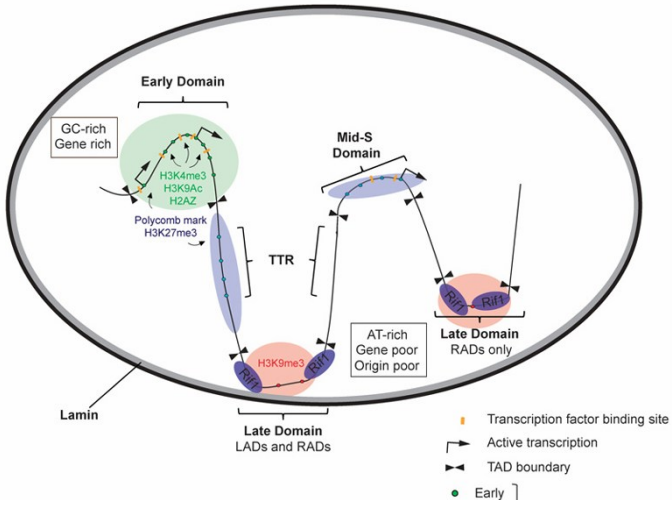

Figure 2.2. Cell replicative domains in nuclei and localization. Early and mid domains are found mostly towards the center of the nucleus (respectively in green and blue), while late replicating domains are located closer to nuclear membrane (in red) because they can be associated with the lamina (LADs) or with Rif1 (RADs). Adapted from Prioleau et MacAlpine.

Based on a technique for origin mapping, called short nascent strand (SNS) enrichment, an initial large-scale analysis over a fraction of the human genome (1%) revealed an enrichment of replication origins in transcriptional regulatory elements such as Transcription Start Sites (TSS) and enhancer regions. CpG islands (CGIs) were identified as the most efficient origins, and about 35% of all origins were of this type (Cadoret et al., 2008). Later, larger scale analysis confirmed these initial findings genome wide and found that G-quadruplexes were particularly abundant within replication origins in mouse and human cells (Cayrou et al., 2012). About 80% of origins overlap with G-quadruplexes. Another genome-wide study detected around 80 000 origins of replication in five human cell lines and showed that not all of the potential G4 sequences (PQSs) overlapped with a replication origin, suggesting that PQSs may be important determinants of origin specification but that they are not sufficient on their own (Picard et al., 2014). Another study using a different origin mapping technique (Ini-Seq2) showed a very high overlap with human SNS-Seq origins (Guilbaud et al., 2022). To summarize many studies of the last two decades, one can distinguish efficient origins which are active in each cell cycle (also sometimes called core origins or constitutive origins), and which are activated in early S-phase, from inefficient origins which do not fire in each cell cycle and each cell, and which are mainly activated in late S-phase. Efficient origins are often located in open

chromatin and in nucleosome-free regions, are associated with active promoters and CpG islands (CGI). However, no strict genetically defined origin consensus sequence has been identified so far in vertebrates.

2.2 G-quadruplexes

Canonical G-quadruplexes are DNA secondary structures found both at replication origins and at telomeres that can form due to the presence of sequences rich in guanine residues, more precisely 4 guanine residues assemble in a planar square linked between each other by Hoogsteen bonds. At least 4 of these planar squares are needed so that they stack one on top of the other and assemble in the final structure; they are connected by loops, so projections of bases connecting two corners of a core (Mukundan and Phan, 2013). The DNA sequences with the ability to potentially form a G4 are the PQSs, putative dimeric G-quadruplex forming sequences (Valton and Prioleau, 2016). They are stabilized by monovalent cations in this order $K^+ > Na^+ \geq NH_4^+ > Li^+$ located in a central position with respect to the tetrad (Jana et al., 2021).

Canonical G4s (Figure 2.3A) can be intramolecular, so formed by 1 single DNA molecule, or intermolecular, thus formed by 2 different DNA molecules. Intramolecular G4s can be parallel, thus all four G-tracts are parallel strands of DNA and the structure contains only propeller loops, meaning loops that link 2 adjacent G-tracts; anti-parallel, where there are 2 parallel and 2 anti-parallel strands, and finally hybrid structures have 3 parallel and one anti-parallel tracts. Loops can therefore be propeller, diagonal or lateral (Figure 2.3B), respectively bridging two diagonally opposite and two adjacent

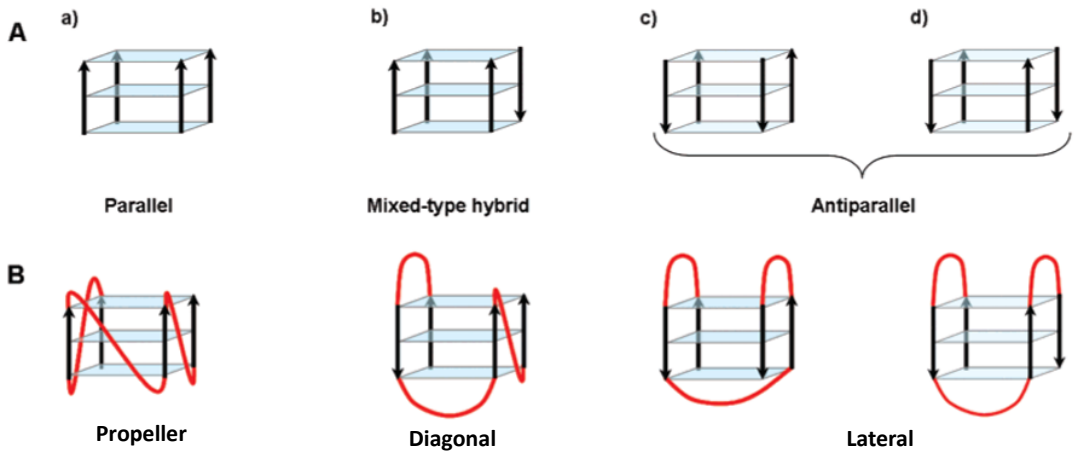

Figure 2.3. G4s canonical conformations and possible loops. A) Possible canonical conformations in G4 structures. B) Different loops that can connect the planar squares, in order propeller, diagonal and lateral loops. Adapted from Bugaut et Balasubramanian.

antiparallel strands (Bugaut and Balasubramanian, 2008), and they can progress either in a clockwise direction or a counter-clockwise direction, meaning that there are a lot of topologies theoretically present, although from a chemical and stoichiometric point of view, only 14 are predicted to exist, and 10 out of these have been described experimentally (Jana et al., 2021). Moreover, loops can vary in length and in nucleotide composition, determining the folded conformation and the thermodynamic stability of the structures, indeed canonical structures contain loops not longer than 7 bases.

These structures can also be non-canonical, for example the G-tracts can be interrupted by other nucleotides therefore generating a bulge, that has the same role as a loop but connects consecutive guanines belonging to the same column. They usually form between the first two residues of the scaffold structure, but they can still form anywhere, and the presence of isolated guanines does not inhibit G4 formation (Mukundan and Phan, 2013). Although we can find many structures containing bulges, they decrease the thermal stability of the structure and depending on their number and positioning melting temperatures change. Moreover, G4s can have one loop of 8-15 nucleotides (4GL15), or they bear a guanine vacancy in the sequence (GVBQ) (Li et al., 2015).

Following the folding rule described by Huppert and Balasubramanian (Huppert and Balasubramanian, 2005), four criteria must be met by a sequence so to consider it able to form a structured canonical G4. First, regarding strand stoichiometry, the formation of interstrand structures is highly disfavored. Second, the possible number of tetrads that can be formed is potentially any, and theoretically the more they are, the higher the stability of the whole structure; single and double G-tetrads have only been reported in a few cases, therefore in the paper they consider only sequences able to form at least 3 tetrads. Third, the stability of the structures is maintained when there are no discontinuities in the G-tracts, so no bases variations. Finally, there must be loops connecting the tetrads but that are also essential to determine the stability and the folding pattern of the G4. The length varies between 1 and 7 nucleotides and with increasing dimension, the stability decreases, although, non-canonical structures can form with a larger number of bases composing the loops. To conclude, the authors define the folding rule as: "a sequence of the form d($G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}$) will fold into a quadruplex under near-physiological conditions", N can be any base, and the physiological conditions are 100 mM KCl and 10 mM Tris–HCl (pH 7.4), therefore we find sequences composed of an alternance of a G-tract of 3 or more guanines and a loop consisting of 1-7 nucleotides (Mukundan and Phan, 2013).

This definition has recently been changed though, because Mukundan and Phan (Mukundan and Phan, 2013) showed the involvement of many isolated guanine

residues and it could be therefore more useful to define G-quadruplex-forming sequences based on the density of guanines present.

Canonical G-quadruplexes are found at promoters of actively transcribed genes, active enhancers, telomeres, topological associated domains (TADs) and at DNA replication origins. This suggests diverse G4 functions at these *cis*-regulatory elements. However, G4s pose also an obstacle for both the replicative and transcriptional machinery. Studies in the chicken DT40 cell line have shown that the delayed replication of one specific PQS positioned on the leading strand template leads to the transient formation of post replicative gaps ahead of the PQS (Schiavone et al., 2014). This results in the uncoupling of DNA replication and the recycling of old histones. Therefore, G4s must be resolved by specialized helicases in order to avoid DNA instability by double strand breaks upon replication fork pausing. These various G4 functions are reflected by the different proteins known to be bound to G4, also called G4-binding proteins (G4BPs), that can have different roles, anchoring, stabilization or unwinding. Sanchez-Martin (Sanchez-Martin, 2023) classifies G4-binding proteins (G4BPs) in factors recruited by G4s that do not affect their structure and that do. The latter category is then subdivided into proteins having a stabilizing and a destabilizing effect on G4s, meaning that they promote the formation of a stable structure, and they unwind it, respectively. Moreover, these proteins can be involved in chromatin remodeling and histone modifications, telomere binding, replication and transcription. What is more of interest for us are the proteins involved in the last two processes mentioned above, for example the BLM and WRN helicases both unwind the double strand but respectively on the leading and lagging strand. Fanconi anemia complementation group J (FANCJ) on the other hand is a helicase that removes G4s already structured to allow efficient replication; DEAH box protein 11 (DDX11) helicase resolves both canonical and non-canonical G4s. Even more interesting is the role recently uncovered for BRCA1, a tumor suppressor but also a G4BP of which the function is still unknown (Brázda et al., 2016). As for G4BPs involved in transcription, we find transcription factors like PARP1, SP1, MAZ, that bind G4s at the level of gene promoters, but also other factors like nucleolin (NCL), that induces CMYC G4 formation, but also nucleophosmin (NPM1) that binds this structure, and TP53 does it as well. Finally among chromatin remodelers and histone modification proteins we find DNMT1 that binds G4s to inhibit the methylation of CpG islands, BRD3 that favors transcription at G4 sites by recognizing hyperacetylated chromatin; many members of the SWI/SNF family like ATRX and the SMARCA proteins; to conclude, recent evidence has been found demonstrating the binding of CTCF to G4s (Samaniego-Castruita et al., 2025).

2.3 How to detect G4s

Among the most common methodologies to identify and map G4s in the genome, we can find G4-seq, G4-miner, G4-ChIP-seq, G4-CUT&Tag, LiveG4ID-seq, Chem-map, G4access and HepG4-seq. I will summarize these techniques based on the review published by Song et al. in 2025 (Song et al., 2025).

G4-seq combines DNA polymerase stalling assays and Illumina sequencing; it relies on the stabilization roles of monovalent cations, therefore G4 formation is induced under Na^+ conditions, unfavorable, and under K^+ conditions, favorable and the two situations are then compared in terms of sequencing read quality.

G4-miner is based on fluctuations in sequencing quality scores during standard Illumina sequencing because G4 unstable structures should cause unexpected variations in these terms but not sequence mismatches and therefore it works in non-stabilizing conditions. These data are then cross-referenced with known G4 mapping data for validation.

G4-ChIP-seq requires specific antibodies or probes that recognize G4s in order to immunoprecipitate the chromatin containing it that is thus sequenced. One possibility is to use the BG4 antibody, developed by the Balasubramanian group, the other is to use the G4P, that I will enter more in detail about in the next paragraph since it's the methodology we use in the lab.

G4-CUT&Tag combines the use of the BG4 antibody with the CUT&Tag technique, so the Tn5 transposase is used to cleave DNA at G4 binding sites so to add sequencing adapters there.

LiveG4ID-seq is a methodology employing live-cell labeling therefore using a live probe, stably expressed in the nucleus under tetracycline induction and that binds G4s with high specificity. The precise localization is allowed by the fusion between a single-chain antibody (specific for G4) and GFP.

Chem-map relies on biotinylated G4 ligands to label the structures in live cells as well, using an antibody against biotin (primary) and a secondary antibody conjugated with the Tn5 transposase again involved in cutting the DNA and inserting adapters for the consequential sequencing.

G4access on the other hand detects G4-forming sequences in open chromatin regions by cutting DNA before G4s through the use of MNase and therefore sequencing the 100 bp fragments obtained.

Finally, HepG4-seq combines G4-hemin complex-mediated proximal biotinylation and CUT&Tag to map G4s *in vivo*. The concept is the same as the biotinylation approach,

because we have the formation of a complex with peroxidase-like activity able to generate phenoxyl radicals that covalently bind to G4 structures and DNA sequences in a radius of 30-60 bp. This complex is formed by the hemin protein, able to mimic peroxidase activity, but also to bind G4s tightly *in vivo* and *in vitro* without affecting its conformation.

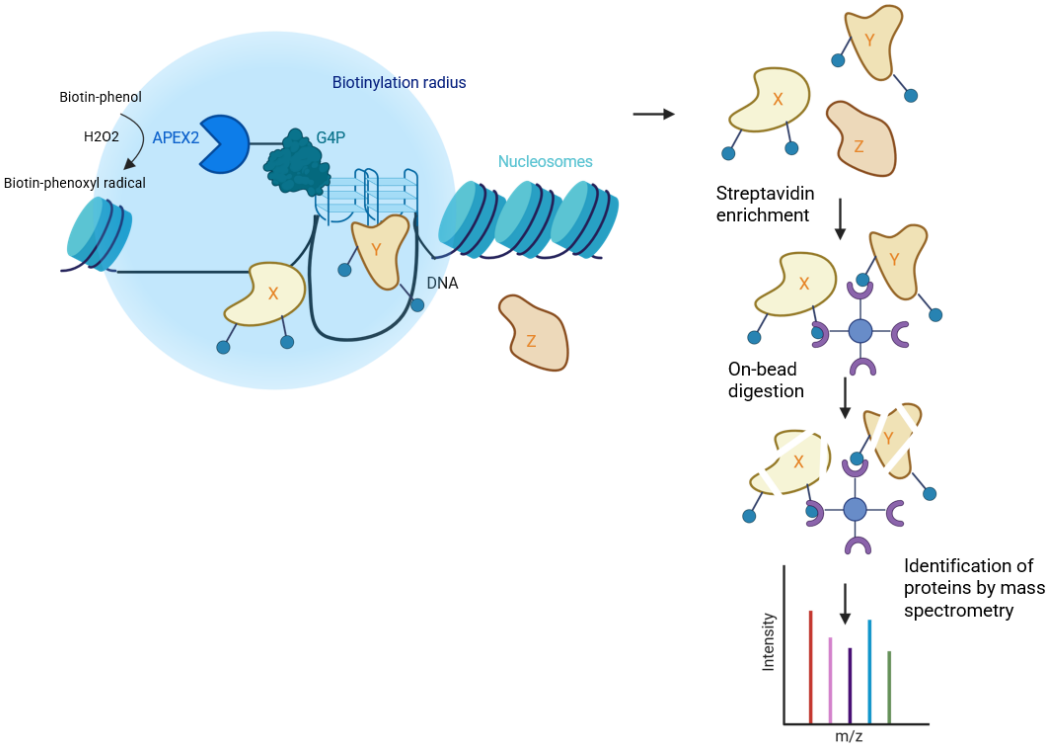

Figure 2.4. Schematic representation of the biotinylation assay. The engineered ascorbic acid peroxidase can generate biotin-phenoxyl radicals that as well bind to amino acid residues in a radius of 10-20 nm. Subsequently through the streptavidin enrichment and the mass spectrometry analysis proteins can be identified. Made with BioRender.

To conclude, I wanted to quickly describe another more general technique that is proximity biotinylation (Figure 2.4), that relies on the ability of the ascorbic acid peroxidase to oxidize, in the presence of H_2O_2, biotin-phenols into biotin-phenoxyl radicals. These molecules have a very short life but are as well able to biotinylate amino acid residues in a radius of approximately 10-20 nm, therefore, after pull down assay using streptavidin the proteins are identified through mass spectrometry. Through this technique, we are able to identify which proteins can be found at the level of PQSs so that are potentially G4 binding proteins. Rhee et al. (Rhee et al., 2013) developed an engineered ascorbate acid peroxidase of 28 kDa, APEX, active in all cellular compartments, and only 2 years later Lam et al. (Lam et al., 2015) generated an evolved version of APEX, termed APEX2, that overcomes the low sensitivity of APEX.

We initially planned to use this methodology during my internship as well, to identify protein interactions happening at the level of the origins of replication and determine whether there is a connection with G4BPs. As a consequence, the plasmid constructs we generated include the APEX2 gene, that transcribes into the protein fused either to the FLAG tag (negative control cell line) or to the G4P-FLAG tag. Nevertheless, we ended up using mass spectrometry for proteins analysis.

Depending on the method, between 10 000-500 000 G4 have been mapped in mammalian cells, illustrating the difficulties to map structured G4s. I will focus here only on the method which uses a recently developed G4-binding probe (G4P).

2.4 G4-Probe-ChIP-seq

A G4 binding probe (G4P) was artificially generated (Zheng et al., 2020) based on the G4-binding sequence of RHAU helicase (RNA helicase associated with AU-rich element), also called DHX36. RHAU is implicated in the regulation of multiple cellular functions including transcription, pre-mRNA splicing, translation, telomere maintenance, genomic stability (Schult and Paeschke, 2021). It is a helicase of the DEAH (Asp-Glu-Ala-His) box family which has been shown to tightly bind and resolve RNA and DNA G4 structures with high specificity (Heddi et al., 2015). The DHX36/RHAU helicase contains a DHX36-specific motif (DSM) at the N-terminal, essential for G4-binding (Figure. 2.5A, B). A high-resolution 3D structure of the helicase demonstrated that the DMS folds into a DNA-binding-induced α-helix and selectively binds a parallel G4 (Chen et al., 2018).

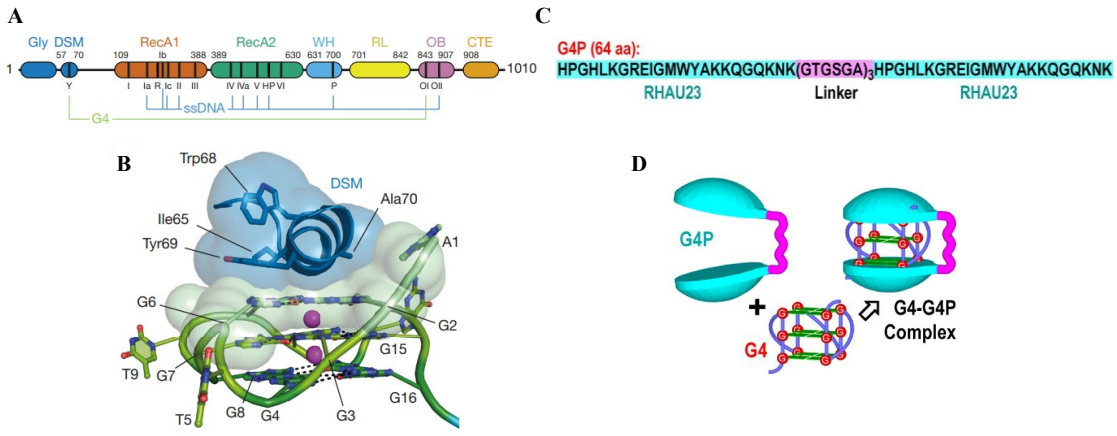

Figure 2.5. RHAU/DHX36 helicase structure and derived G4-Probe (G4P). A) DHX36 domain organization with the DHX36-specific motif (DSM). B) Crystal-Structure of DSM of DHX36 bound to a G4. Adapted from Chen et al. 2018 C) Amino acids sequence of G4P composed of two DSM elements (RHAU 23 aa) and a flexible linker. D) Anticipated clamping binding to G4s. Adapted from Zheng et al. 2021.

The 64 amino acids long G4 probe (G4P) is composed of two modules of the G4-binding sequence of the DSM (RHAU 23) of the helicase and is separated by a flexible linker (12 aa) (Figure 2.5C, D). They found that the synergy between two G4 binding domains strongly improved affinity and selectivity towards G4s and demonstrated its ability to bind other canonical and non-canonical structures in a low nanomolar range. As shown in Figure 2.5B the crystallized DSM/G4 structure is consistent with the anticipated clamping hypothesized by Zheng et al., therefore the two 23 amino acids domains of RHAU possibly form α-helices that enclose the G4.

After transient transfection or stable integration (knock-in) in human A549 cells and in chicken DF-1 cells, Zheng et al. demonstrated G4P binding capacities both *in vitro* and *in vivo* to G4-DNA. By performing G4P-ChIP-seq experiments they detected 123 274 G4P peaks which are enriched at canonical and non-canonical G4s, at active promoters and co-localized with several known G4BPs in human cells. The advantage of the G4P-Seq method is the *in vivo* binding of the endogenously expressed G4 peptide. Compared to G4-Seq using an anti-G4 antibody, this avoids potential artefacts of crosslinking of chromatin prior to anti-G4 antibody binding. In addition, the simple composition of the G4P could minimize potential non-specific interactions with other proteins.

2.5 Clustered G4s and DNA replication origins are highly enriched at G4P-peaks

As mentioned, above, putative G-Quadruplex forming Sequences (PQSs) have been functionally associated with a subset of replication origins. Studying a model replication origin in DT40 chicken cells, the Prioleau lab has shown the critical importance of PQSs. More recently, they found that a dimeric PQS (i.e., two PQSs on the same strand and separated by at most 100 bp) forms a minimal combination of *cis*-elements (minimal origin) sufficient for replication initiation (Poulet-Benedetti et al., 2023). They showed that 52% of clustered PQSs in chicken are associated with origins of replication compared to 31% monomeric PQSs genome wide. Clustered PQSs contain 2-6 G4s inside 100bp on one strand and constitute around 13% of all G4s are in clusters in the chicken genome. In addition to the presence of these clustered G4s, efficient origins in vertebrates colocalize with nucleosome-free regions (NFRs) and the histone variant H2A.Z, which has been shown to promote origin activity. Thus, a dimeric PQSs or more general clustered PQSs constitutes a new nucleic acid feature associated with one-third of strong replication initiation sites genome wide in human and chicken cells (Figure 2.6A).

The Prioleau lab also compared the G4P-IP peaks from Zheng study with their genome-wide origin maps (SNS-Seq data) in human and chicken cells (Picard et al., 2014) (Massip et al., 2019). They found that clustered PQSs containing an origin are also more associated with structured G4s detected by G4P-IP than clustered PQSs without origin. The association of monomeric PQS with G4P peaks is lower, meaning that the

probability of finding a structured G4 is higher in genomic windows containing clustered PQSs (Figure 2.6B).

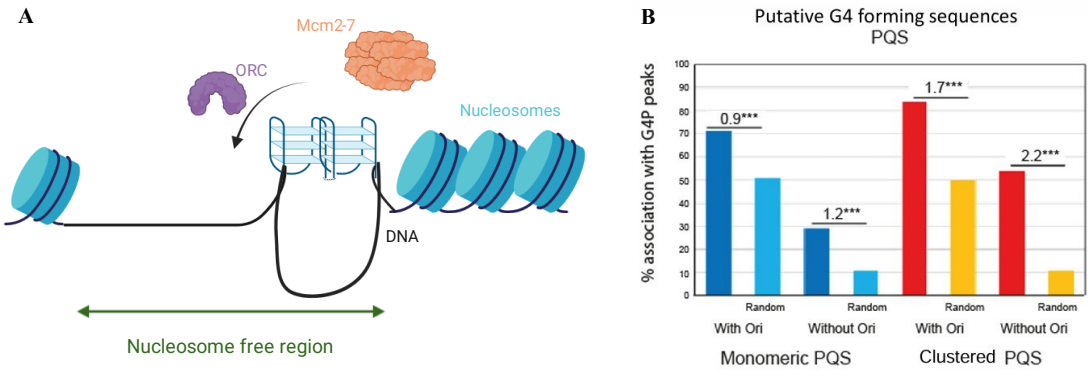

Figure 2.6. Efficient replication origins are associated with clustered G4. A) Model of minimal origin features with a dimeric PQS at the border of a nucleosome free region. ORC complex can bind to the NFR and load MCM helicase in G1 phase. B) Association of G4P peaks with monomeric and clustered PQS and replication origins in DT40 chicken cells. Adapted from Poulet-Benedetti et al.

2.6 Objectives of my internship project

It is unclear how clustered G4s contribute exactly to the constitution of an efficient replication origin. It has been suggested that clustered G4s tend to exclude nucleosomes, thereby potentially favoring pre-RC formation (Fenouil et al., 2012). Alternatively, G4 formation may facilitate DNA unwinding and, therefore, the initiation of replication. Finally, structured G4s may be recognized by specific factors involved in the formation of a functional origin. For example, Rif1 and MTBP, both regulators of the initiation of DNA replication, have been shown to bind G4s *in vitro* (Kobayashi et al., 2019; Kumagai and Dunphy, 2020), respectively.

The lab has shown that dimeric PQSs are at the boundary between nucleosomes and NFRs. To date no potential associated *trans*-factors at PQSs, such as chromatin remodelers, have been involved in the establishment of active origins. To determine whether a specific set of proteins is present at these origins, the lab plans to perform an unbiased proteomic approach using proximity biotinylation with the ascorbic peroxidase.

Since the G4P-peaks show a good overlap with efficient origins, the Prioleau lab decided to use the G4P as bait. They constructed a stable DT40 cell line expressing a G4P-APEX2 fusion under a doxycycline inducible promoter using the Tet-ON 3G system (Figure 2.7A). A Flag-tag is present at the N-terminal of G4P for the use of chromatin immunoprecipitation. The initial G4P expression conditions had been determined before my arrival in the lab. Further on, they found that short induced expression of G4P-APEX2 does not affect the cell cycle progression, thus implying

that the induced and controlled expression of this G4 binding probe seems not toxic to cells or does induce major DNA damage. However, a control cell line necessary for the proximity labelling approach with the APEX2 alone under the TRE3G promoter had not been obtained yet.

The aim of my internship was to better explore the use of the artificial G4P fused to APEX2 to 1. detect structured G4s in the genome and 2. describe the proteome around these structures in chicken DT40 cells. To verify that the G4P fusion with APEX2 (28kDa) does not alter the binding specificity of the G4P as described above, I will first perform chromatin immunoprecipitation against the flag-tagged G4P-APEX2 after expression induction (G4P-ChIP) followed by qPCR and high-through put sequencing to verify the presence of G4P peaks in correspondence to PQSs and mapped replication origins in DT40 cells. Second, for the proteomic approach using proximity biotinylation with APEX2, I needed to obtain DT40 control cell line with an inducible APEX2.

While this cell line was under construction, we decided to explore an alternative proteomic approach using G4P-ChIP (Figure 2.7B).

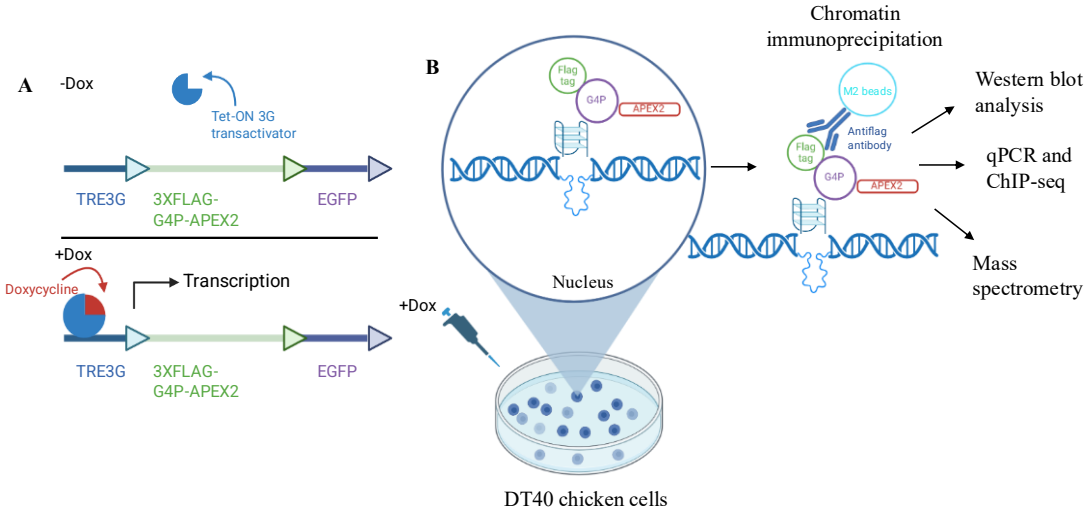

Figure 2.7. Schematic presentation of G4 probe (G4P)-APEX2 expression system in DT40 cells and methodologies used in my study. A) G4P-APEX2 expression using the TetON induction system; doxycycline induction changes the conformation of the transactivator so it binds to the promoter and activates Flag-G4P-APEX2 transcription. B) Upon doxycycline induction, FLAG-G4P-APEX2 protein binds to structured G4s in the nucleus. Chromatin immunoprecipitation with an anti-Flag antibody should detect structured G4s by qPCR, ChIP-seq and proteins bound to those G4 by mass spectrometry.

3 Materials & Methods

3.1 Model system and cell culture conditions

Our biological model system is the DT40 chicken cell line. The chicken genome is composed of 1.2×10^9 bases and 2n=80 chromosomes. DT40s have several advantageous characteristics, including a relatively short cell cycle (8 hours), a stable karyotype, and, most importantly, highly efficient homologous recombination (HR). They derive from the generation of bursal lymphoma with avian leukosis virus transplanted twice *in vivo* and thus cultured. WT and genetically modified DT40 chicken cells are cultured in RPMI 1640 medium supplemented with Glutamax (Thermo Fisher Scientific #61870010), containing 10% fetal bovine serum, 1% chicken serum, 0.1mM β-mercaptoethanol, 200U/mL penicillin, 200µg/mL streptomycin, and 1.75µg/mL of amphotericin B, at 37°C, and in a 5% CO2 condition. Given their fast growth rate, they must be kept in a concentration between 0.2 and 2.5 million/mL when plated daily. The cells in which the plasmids described below were stably inserted by homologous recombination were selected in a medium containing a final concentration of 1 µg/mL puromycin.

3.2 Plasmid construction and cell line generation

The FLAG-G4P-APEX2 plasmid (Figure 3.1) was previously constructed in the lab, as well as the DT40 cell line clones, which stably integrated this construct in an open chromatin region of chromosome 1 by homologous recombination (chr1:91742981+91744494, GalGal5). The expression cassette contains a sequence of a flag tag, the *g4p* and *apex2* and that is transcribed under the control of the TRE3G promoter containing the tetO operators (7 repeats of 19 bp). This promoter is activated

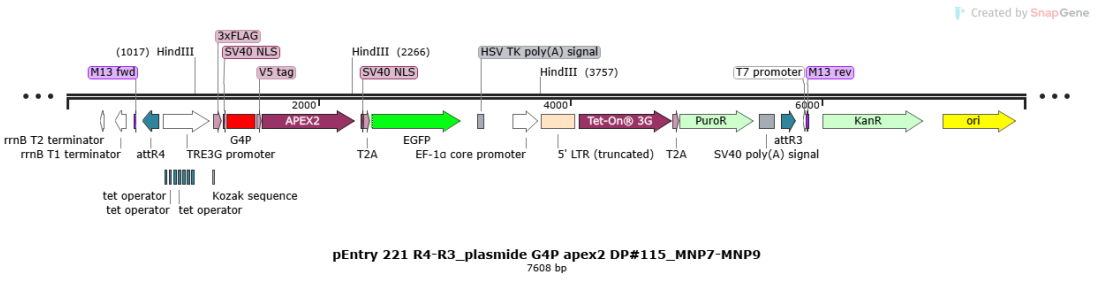

Figure 3.1. FLAG-G4P-APEX2 plasmid. The plasmid previously generated in the lab contains a fusion of the flag tag sequence, the *g4p* sequence, the *apex2* gene and the *egfp* gene under the control of the TRE3G promoter and the Tet-On 3G transactivator protein. EGFP is translated separately from the rest of the construct due to the T2A sequence. Puromycin resistance gene is used for clone selection.

by the Tet-On 3G protein, a transactivator constitutively transcribed but that can bind the promoter only when undergoing a conformational change in the presence of doxycycline (Dox); therefore, the Flag-tagged G4P-APEX2 fusion protein is expressed upon induction with doxycycline. Under the control of the same promoter, we also find a reporter gene, enhanced green fluorescent protein (*egfp*), that will not be fused in the structured protein due to the presence of the T2A ribosome skipping sequence. Finally, we find an antibiotic cassette for puromycin for selection. The Flag-tagged G4P-APEX2 protein has a size of 46 kDa. The FLAG-APEX2 plasmid (Figure 3.2) was also previously constructed by the lab, but DT40 clones for inducible expression of the *apex2* had not been obtained. The plasmid presents the same characteristics as the one described above, but the *g4p* sequence is omitted. The resulting protein has a size of 28 kDa.

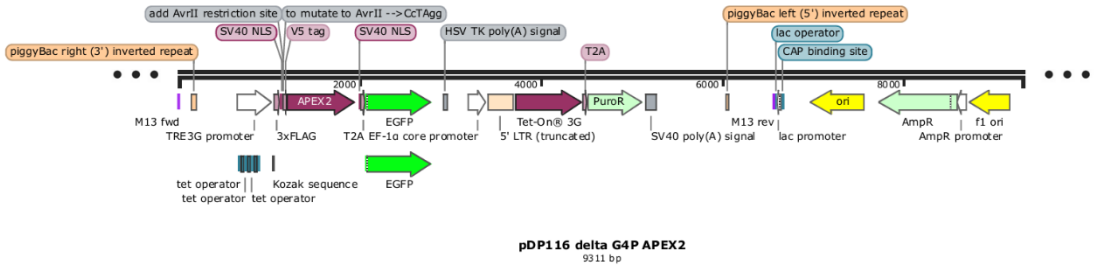

Figure 3.2. FLAG-APEX2 plasmid. The plasmid previously generated in the lab contains a fusion of the flag tag sequence, the *apex2* gene and the *egfp* gene under the control of the TRE3G promoter and the Tet-On 3G transactivator protein. EGFP is translated separately from the rest of the construct due to the T2A sequence. Puromycin resistance gene is used for clone selection.

To generate the modified DT40 cell clones, we linearized the plasmid with the ScaI HF restriction enzyme that does not cut inside the expression cassette. WT DT40 chicken cells were electroporated at 550 V and 25 µF. A limiting dilution is carried out to obtain one cell per well in 96-well plates. Cells were then grown in the presence of puromycin. After ten days, single clones are isolated and transferred to 24-well plates and then to 6-well plates to grow. Clones with a better growth rate were selected to be induced with doxycycline at 1 µg/ml for 21h and to be analyzed for their EGFP expression through FACS.

3.3 SDS-PAGE and Western blotting

To separate proteins by size on sodium dodecyl-sulfate polyacrylamide gel electrophoresis (SDS PAGE), I used the Mini-PROTEAN TGX stain-free gels 4-20% (Bio-Rad) run at 15-20 mA in the Tris-Glycine SDS 1X migration buffer (25 mM Tris, 192 mM glycine, 0.1% SDS). The migrated proteins are then transferred to a polyvinylidene fluoride (PVDF) membrane (Millipore, 0.45 µm) in Bjerrum Schafer-

Nielsen transfer buffer (48 mM Tris, 39 mM glycine, 20% isopropanol, 0.037% SDS) at 1 mA/cm^2. To detect the expression of the G4P-APEX2 protein, I used as a primary antibody the M2 anti-flag antibody (Sigma-Aldrich, F1804), at 1:1000 dilution factor in 5% milk powder in 1X TBS 0.1% Tween, followed by a secondary antibody Horse Radish Peroxidase (HRP)-anti-mouse (Bethyl Lab, A90-516P) (1:10 000). For the normalization of G4P APEX2 and APEX2 bands in cellular extracts, we stripped antibodies off membranes and incubated with an anti-tubulin mouse antibody (Sigma-Aldrich, T5168) in a 1:2000 dilution in 5% fat-free milk powder in 1X TBS 0.1% Tween. For detection of G4P-APEX2 immunoprecipitated from crosslinked chromatin, beads were eluted with 2x Laemmli buffer (containing 4% SDS) for either 5' at 30°C, then 40' at 95°, or directly 40' at 95°C. Histone H3 was visualized for chromatin loading control using an anti-H3 rabbit antibody (Bethyl Lab, A300-823A) at 1:2000 dilution, followed by a secondary antibody Horse Radish Peroxidase (HRP)-anti-rabbit (Cell signaling, 7074) (1:10000). Specific proteins bands are detected by chemiluminescence with the ECL Select Western Blotting Detection Reagent (Cytiva) on a Chemidoc Imager (BioRad). For quantification of western blot bands, I used the Image Lab software (BioRad).

3.4 Chromatin immunoprecipitation

The cells were either treated or not with doxycycline (10 ng/ml) for 16 hours. The chromatin was prepared for the chromatin immunoprecipitation (ChIP) using the truChIP shearing kit (Covaris, PN 520154), following the protocol for "high cell condition" with 30x10^6 cells. Briefly, cells were fixed in 11.1% formaldehyde for 10 minutes, then quenched. Cells were lysed, nuclei were isolated, and chromatin was sonicated for 20 minutes using a M220 Covaris sonicator (conditions are: PIP 75, duty factor 10%, CPB 200, setpoint temperature 7°C, range of temperatures 4/10°C, water level full, maximum cell number 15 million and sample volume 1 mL). Purified chromatin samples in 1X shearing D3 buffer (1mM EDTA, 10 mM Tris-HCl pH 7.6, 0.1% SDS), supplemented with 10% glycerol, were stored at -80°C. G4P-APEX2 chromatin immunoprecipitation (ChIP) was performed with anti-Flag M2 magnetic beads (Millipore, M8823) that were incubated overnight with the chromatin. I used 500 µl of chromatin (corresponding to 7.5x10^6 cells) per 10 µl beads; the supernatant was removed, and the beads were then washed 4 times with 1X IP buffer (5 mM Tris-HCl pH 8, 0.5 mM EDTA, 150 mM NaCl, 1% Triton). Bound DNA was eluted with elution buffer (50 mM Tris-HCl pH 8, 10 mM EDTA, 1% SDS) overnight (ON) at 37°C, proteins digested with 100µg/µl of proteinase K at 65°C, ON in agitation at 1000 rpm, and DNA was extracted and ethanol precipitated. For protein analysis, beads were either eluted with 2X Laemmli buffer, followed by SDS-PAGE and western blotting, or beads were digested with trypsin for MS-analysis (see section 3.8).

3.5 DNA extraction and precipitation

DNA was purified by Phenol/Chloroform extraction using the MaxTract high-density phase lock gel tubes (Quiagen, 129056), precipitated in ethanol, and resuspended in 1X TE buffer (pH 7.4). The DNA was then quantified with the Qubit assay kit (Thermo Fisher Scientific), broad range or high sensitivity.

3.6 G4P-qPCR

Purified DNA after ChIP -/+ Dox and Input DNA before ChIP -/+ Dox were analyzed in duplicates by qPCR using a Roche LightCycler 480 and 480 Sybr Green Master I kit (Roche, 04887352001). Two different regions were analyzed with two specific primer pairs (Table 3.1). PCR reactions were done in 384-well plates, adding 2 µl of DNA sample and 13 µl of mix solution, composed of: 1 µl forward primer at 10µM, 1 µl reverse primer at 10µM, 7.5 µl of Sybr Green Mix 2X concentrated, and 3.5 µl of sterile water. I included sonicated genomic DNA of known concentrations between 2 ng/µl and 0.00315 ng/µl to obtain the standard curve for absolute quantification. The samples are diluted to a concentration ranging from 2 ng/µl to 0.00315 ng/µl. The qPCR program was as follows: initial denaturation 95°C 5 min; 40 cycles: 95°C 10 sec, 61°C 20 sec, 72°C 20 sec; melting curve: 95°C 5 sec, 70°C 1 min, 70°C to 95°C; cooling: 40°C 30 min.

Primer	Sequence	Description
β-actin forward	AAACCGGCCTTGCACATACC	They amplify a sequence of 90 bp
β-actin reverse	TCCCTTCTCTGTTCCTCCGC	They amplify a sequence of 90 bp
Cond-1 forward	TTGGTGCAGTGCCTCAGATAG	They amplify a sequence of 107 bp
Cond-1 reverse	ATGTCGCTTGTCACGATGGAT	They amplify a sequence of 107 bp

Table 3.1. Sequences of the primers used for the qPCR amplification.

3.7 G4P-ChIP-seq

Samples were sonicated before library preparation to obtain a mean size of around 250 bp (Covaris M220 sonicator: PIP 30, duty factor 20%, cycles per burst 50, temperature 20°C, time 80 seconds, and volume 20 µl). For the sequencing steps, I prepared the library using the NEBNext Ultra II DNA Library Prep Kit for Illumina, following the manufacturer's protocol with same amount of DNA for all samples (50 ng): Input -Dox/+Dox, ChIP -Dox/+Dox. I use the SPRIselect Beads (Beckmann Coulter, B23317)

for the purification steps. The samples prepared for the sequencing are checked through an Agilent bioanalyzer chip before the sequencing to determine concentration and size distribution. Samples were sent to the GENOM'IC sequencing platform at Cochin Institute for 2X50 bp paired-end sequencing.

3.8 Mass Spectroscopy of G4P-ChIP proteins and pathway enrichment analysis

Samples for mass spectroscopy analysis at the ProteoSeine platform at the IJM were prepared as follows: beads from the ChIP pulldown assay were incubated overnight at 37°C with 20 µL of 50 mM NH_4HCO_3 buffer containing 1 µg of sequencing-grade trypsin/Lys C mix. Before LC-MS/MS analysis, the digested peptides were loaded and desalted on evotips provided by Evosep (Odense, Denmark) according to the manufacturer's procedure. Samples were analyzed on a timsTOF Pro 2 mass spectrometer (Bruker Daltonics, Bremen, Germany) coupled to an Evosep one system (Evosep, Odense, Denmark) operating with the 40SPD Whisper Zoom method developed by the manufacturer. MS raw files were processed using Spectronaut version 19.4.241104.62635. Data was searched against the Gallus Chicken UniProt proteome (UP000000539) with the addition of the bait G4P-APEX2 sequence. Quantification was performed using the Spectronaut Quantification Module, with all default parameters. Proteins were inferred using the automatic features of Spectronaut, with the algorithm IDPicker. For pathway and process enrichment analysis, I used Metascape (metascape.org) for my enriched list of proteins (>1.5). Pathway and process enrichment analysis have been carried out with the following ontology sources: KEGG Pathway, GO Biological Processes, Reactome Gene Sets, Canonical Pathways, CORUM, and WikiPathways. All genes in the genome have been used as enrichment background. Terms with a p-value < 0.01, a minimum count of 3, and an enrichment factor > 1.5 (the enrichment factor is the ratio between the observed counts and the counts expected by chance) are collected and grouped into clusters based on their membership similarities. More specifically, p-values are calculated based on the cumulative hypergeometric distribution. The most statistically significant term within a cluster is chosen to represent the cluster.

3.9 Detection of G4P expression in DT40 cells by immunofluorescent microscopy

DT40 cells grow in suspension. For immunofluorescence studies, $0{,}5x10^6$ cells in culture medium were centrifuged at 1200 rpm for 5 minutes on poly-L-lysine-coated coverslips. The medium was taken off, and the cells were fixed with 4% paraformaldehyde/1X PBS for 5 minutes. The cells were blocked for 30 min in 1X PBS, 0.02% SDS, 0.1% Triton, 2% BSA. Then, cells were incubated with the anti-Flag mouse monoclonal antibody (F-1804, Sigma-Aldrich) for one hour at 37°C in a humified chamber (1:200 dilution in PBS 1X, 0.02% SDS, 0.1% Triton, 2% BSA). After washing 4 times 10 min with 1X PBS, 0.02% SDS, 0.1% Triton, the anti-mouse

AlexaFluor 564 antibody (Thermo Fisher Scientific, A21123) (1:200 dilution in 1X PBS S/T + BSA) was added for 30 min, 37°C, together with DAPI (stock 0.5 mg/ml) at 1:100 dilution for DNA detection. Coverslips were washed as before and mounted in mounting medium (citifluor, science services). Images were taken with a 100x objective using a Zeiss Spinning Disk CSU-X1 confocal microscope and further treated with ImageJ software.

4 Results

4.1 Inducible expression of G4P-APEX2 in DT40 cells

An expression cassette of the artificial G4-Probe (G4P) fused to the APEX2 in the C-terminus and a 3xFlag-tag sequence in the N-terminus had been inserted by homologous recombination into the Chr1 of DT40 chicken cells and a stable cell line had been selected (clone 1), before my arrival in the lab. Upon induction with Doxycycline (10 ng/ml) the Tet-On3G transactivator undergoes a conformational change and is able to bind to the TRE3G promoter leading to the expression of the G4P-fusion protein. After 16h of induction, western blot analysis of cellular extracts using an anti-Flag antibody detected a specific band migrating between 37 and 50 kDa, which showed the predicted expression of the 46 kDa Flag-G4P-APEX2 fusion protein (Figure 4.1). As a loading control we used tubulin. No specific band was observed in the absence of doxycycline, confirming the tight regulation of this expression system. No further increase in the G4P-APEX2 protein was detected after 24h of induction. We therefore chose to use 16h induction by doxycycline for the following experiments.

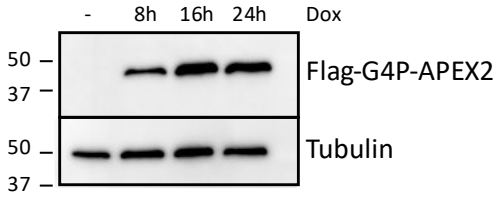

Figure 4.1. Tight regulation of G4P-APEX2 expression in DT40 cells. G4P-APEX2 expression was induced after indicated times of doxycycline addition, cellular proteins were separated by SDS-PAGE and Western blot analysis was performed using an anti-Flag antibody. Tubulin is used as a loading control. Proteins dimensions are indicated in kDa.

4.2 Chromatin immunoprecipitation (ChIP) of Flag-tagged G4P-APEX2 in DT40 cells

For the ChIP, I prepared chromatin from DT40-Flag-G4P-APEX2 cells -/+ Dox using the TruChIP kit as described in Materials&Methods according to lab protocols. Briefly, cells were incubated with formaldehyde for 10 min to crosslink proteins bound to chromatin, nuclei were isolated, and chromatin was sonicated to obtain DNA fragments between 250-500 bp. Chromatin was incubated with anti-Flag-M2 magnetic beads over night at 4°C in IP-Buffer and beads were washed extensively to reduce non-specific binding of chromatin to magnetic beads. Immunoprecipitated proteins were eluted at 30°C with 2x Laemmli buffer in non-reducing conditions to avoid co-migration of the heavy IgG chains. The different samples along the experiment were analyzed by

western blotting using anti-Flag antibody (Figure 4.2) (and Histone H3 as chromatin control). I observed a specific band of G4P-APEX2 in chromatin samples before incubation with M2 beads indicating that G4P-APEX2 can be found on crosslinked chromatin, thus most likely binds to chromatin. Furthermore, a specific G4P-APEX2 band was also detected in the immunoprecipitated fraction. Small amounts of non-bound G4P-APEX2 were observed in the IP-supernatant. No specific bands were detected, as shown above in the absence of doxycycline.

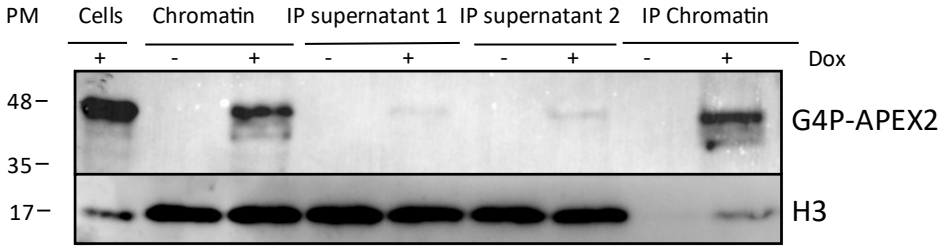

Figure 4.2. G4P-APEX2 binds to chromatin and can be immunoprecipitated. Chromatin was prepared from DT40-G4P-APEX2 cells in the presence and absence of doxycycline. Western blot analysis using anti-Flag antibody of cell extracts, isolated chromatin, IP supernatant and IP with anti-Flag M2 magnetic beads in the presence and absence of doxycycline. Proteins dimensions are indicated in kDa.

4.3 Expressed G4P-APEX2 is mainly localized to the nucleus

To check the cellular localization of expressed G4P-APEX2 I performed immunofluorescent staining in DT40 cells containing the G4P-APEX2 construct in the absence and presence of Doxycycline (Figure 4.3). In the -Dox condition there is nearly no signal with the anti-FLAG antibody, besides some background staining due to either the primary or the secondary antibody, spread all over the cell. After induction of G4P-APEX2 expression, there is a strong nuclear staining with the anti-FLAG antibody, confirming the G4P overexpression works under Dox induction, consistent with the western blot analysis. Moreover, since the construct contains a nuclear localization signal (NLS), the signal is nearly exclusively localized in the nucleus, as expected.

I also observed a strong variability of G4P-APEX2 staining patterns, ranging from no signal, whole-nuclei staining or partial nuclear staining, clearly confined to some sub-nuclear structures. We hypothesize that this sub-nuclear localization could correspond to the nucleolus. However, no clear nucleolar foci were detected.

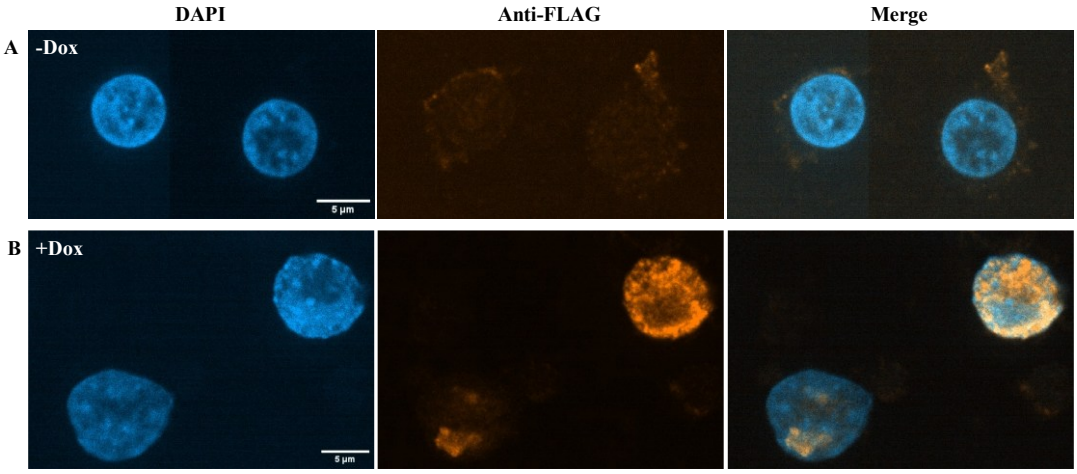

Figure 4.3. Immunofluorescence microscopy of DT40 cells expressing G4P-APEX2. A) -Dox, non-induced cells B) +Dox, induced cells. DAPI was used to stain DNA (blue), anti-Flag antibody followed by an anti-mouse AF568 to stain G4P-APEX2. Photos were taken with a 100x objective using a Zeiss Spinning disk confocal microscope.

4.4 G4P- ChIP qPCR

Next, I purified DNA bound to anti-Flag-M2-magnetic beads after ChIP from uninduced and induced DT40 G4P-APEX2 cells and quantified the DNA by a high-sensitivity Qbit kit. I observe a 4.2-fold increase of the amount of isolated DNA after G4P-APEX2 induction compared to the un-induced control, which showed that we can purify and enrich total DNA after G4P-IP (Figure 4.4A). To determine whether G4P-APEX2 fusion protein specifically binds to G4s in DT40 cells, I performed qPCR analysis, using primers targeting the β-Actin locus, a region in Chr14 containing a large CpG island, PQSs and a replication initiation zone obtained by SNS-Seq (Massip et al., 2019) (Poulet-Benedetti et al., 2023). As a negative control region (Cond-1) we chose primers targeting a region on Chr1 containing neither a CpG island, PQSs nor a replication origin (for more details see Figure 4.7 below). By qPCR, we expected to find an enrichment at the β-Actin locus but not at the Cond-1 locus after G4P-APEX2 expression. I purified DNA after ChIP from uninduced and induced DT40 G4P-APEX2 cells as well as DNA from chromatin before IP (Input DNA) and performed qPCR reactions in duplicates using a Lightcycler 480. Absolute DNA was quantified using a standard curve using sonicated genomic DT40 DNA. For normalization, ratios of ChIP/Input DNA were calculated for both loci and compared in the absence and presence of doxycycline. Two independent ChIP experiments followed by qPCR analysis were performed (Fig 4.4B, C).

I could observe a 7.6-fold and a 5-fold enrichment, respectively, after doxycycline addition at the G4 containing β-Actin locus as expected. However, I observed also

some, although less important, enrichment at the Cond-1 control locus after G4P induction. Altogether, I only found a mean 1.6-fold difference, between the β-Actin locus and the control Cond-1 locus upon G4P expression. This unexpectedly low difference between the G4 containing region and the control region by qPCR analysis suggested a relatively high non-specific background of the ChIP signal in non-G4 containing regions. Alternatively, it could also reflect the fact that the qPCR primers were not exactly on the PQSs in the β-Actin locus. Indeed, no other primer pairs could be designed to work in qPCR experiments inside the GC rich regions more downstream of the β-Actin primer pair. To circumvent this problem, we decided to analyze genome-wide the enrichment of G4P immunoprecipitated DNA by high-through-put sequencing.

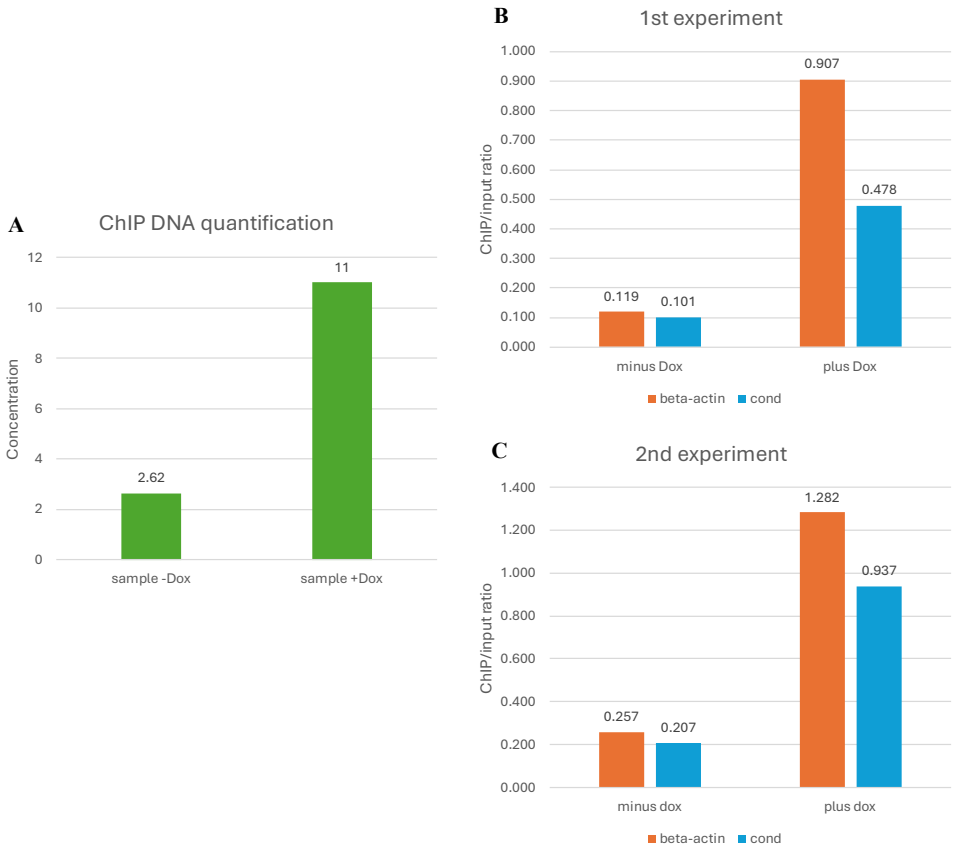

Figure 4.4. G4P ChIP qPCR reveals only a modest increase between a PQS rich locus versus a PQSs poor locus. A) Total DNA after ChIP was quantified by HS-Qbit. B) Quantification of ChIP/Input DNA ratios at a PQSs-rich locus, β-Actin and a PQSs poor locus, Cond-1, in the absence and presence of doxycycline by qPCR, replicate 1. C) Replicate 2.

4.5 Preliminary G4P-ChIP-seq results

Library Preparation

For library preparation, 50 ng of chromatin per sample in absence and presence of doxycycline were prepared. Input DNA and DNA after chromatin precipitation with anti-Flag M2 magnetic beads was purified as described before. DNA was re-sonicated to obtain DNA fragments around 200 bp. Libraries were constructed using the NEBNext Ultra DNA library Kit from Illumina according to the manufacturer's instructions. DNA concentration and size were determined using Agilent Chip technology (Figure 4.5 A, B). The fragment size distributions and the concentrations were in the range recommended by the GENOM'IC facility at the Cochin Institute.

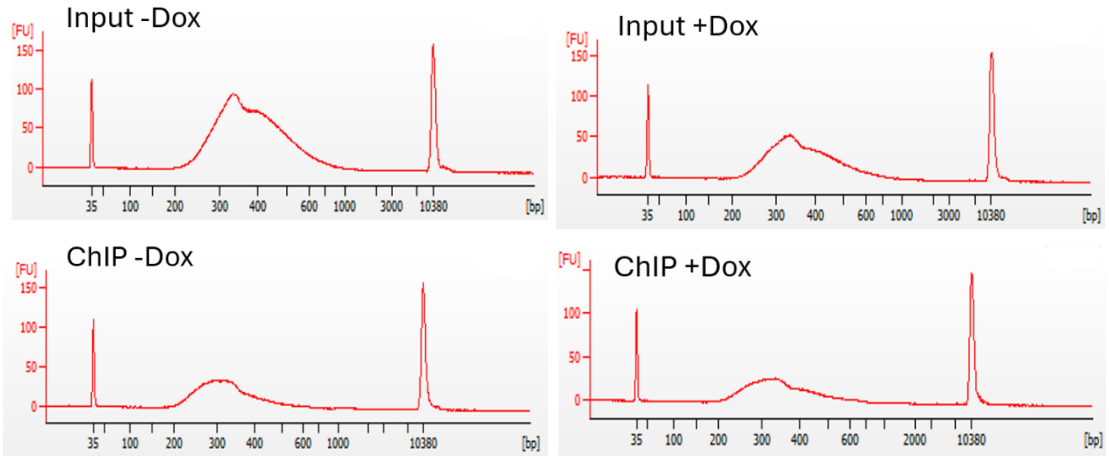

Figure 4.5. Dimension spectrum of the samples analyzed through the agilent chip.

	From [bp]	To [bp]	Corr. Area	% of Total	Average Size [bp]	Size distribution in CV [%]
Input -Dox	200	1000	1994.2	98	397	27.2
Input +Dox	200	1000	1064.4	98	383	27.3
ChIP -Dox	200	1000	670.7	98	336	25.2
ChIP +Dox	200	1000	549.8	98	360	26.8

	Conc. [pg/µl]	Molarity [pmol/l]	Total DNA [ng]	Adjusted molarity [pmol/l]	Volume [µl]
Input -Dox	1451.18	6046.9	110.28968	24187.6	19
Input +Dox	757.49	3265.7	57.56924	13062.8	19
ChIP -Dox	495.26	2393.7	37.63976	9574.8	19
ChIP +Dox	428.77	1961.6	32.58652	7846.4	19

Table 4.1. Chip agilent analysis. Raw data from the chip agilent analysis, each column showing respectively: the range of dimension of the reads; the graphs correlation areas; the percentage of the area occupied by the reads in that range of dimension; the average size of the reads; the size distribution; the DNA concentration; the molarity; the total amount of DNA in the final volume; the molarity in the final volume; the final volume.

Paired-end sequencing was performed for Input -Dox/+Dox and ChIP -Dox/+Dox libraries using a NextSeq500 System from Illumina. Between $80x10^6$ and $110x10^6$ of reads were obtained.

Data processing and preliminary results of G4P-ChIP-Seq

Clean paired-end sequencing data in fastq format were mapped to the chicken genome (GalGal5) using Bowtie2 software using lab internal pipelines giving good alignments (overall alignments between 97.5% and 97.8%). I then imported mapped reads written in Bamformat into Galaxy and calculated reads enrichments using Deeptools (bamcoverage) using 250 bp windows to produce bigwig coverage files. These were further processed to produce files of ChIP/Input in ratio mode (log2 ratio) and normalized with RPKM, using Bigwigcompare.

I visualized the bigwig compare Chip/Input files -/+ Dox in the USCS genome browser together with G4 peaks (Poulet-Benedetti et al., 2023), CpG islands and mapped origins of replication by SNS-Seq (Massip et al., 2019). In general, I observed many peaks in the +Dox condition, but also in the -Dox condition all over the genome suggesting a relatively low signal to noise ratio. Nevertheless, closer inspection focusing on CpG islands often revealed read enrichments at these regions in the +Dox condition compared to the -Dox condition as shown in two regions of interest on Chr10 (Figure 4.6A, B). These peaks are also often correlated with the presence of replication origins or initiation zones. Further bioinformatic analysis is now needed to quantitatively describe the G4P ChIP-Seq data by an adapted peak finding method. Results could then be compared with the published G4P Chip-Seq data set.

We also visualized Chip/Input ratios the two loci analyzed by qPCR, the β-Actin and Cond-1 locus (Figure 4.7A, B). I observed a relatively low enrichment at the qPCR amplicon for the β-Actin locus, but a higher enrichment more downstream. At the Cond-1 replicon, we also observe a small enrichment of Chip/Input ratios in the +Dox condition but none in the -Dox condition. These observations seem to confirm the low enrichment values I obtained by qPCR analysis for these chosen loci.

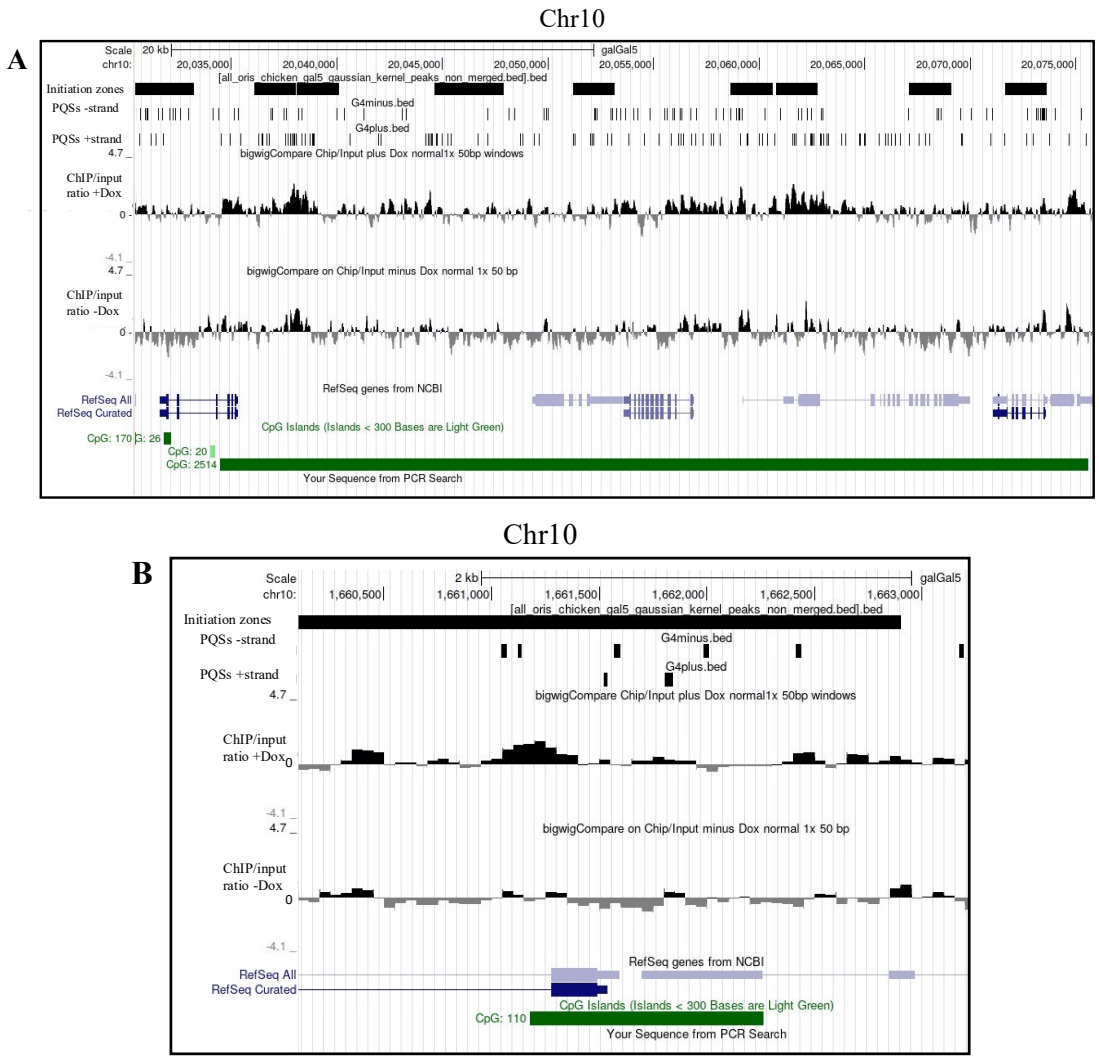

Figure 4.6. Visualization of Chip/Input ratios -/+ Dox in USCS browser for two different regions at Chr10 centered at CpG islands. A) Large window showing the multitude of peaks in + Dox, compared to less peaks with lower amplitude in the -Dox condition. B) Smaller window focusing on one peak showing clear enrichment in + Dox condition. Chip/Input ratios are log2 ratios in 250 bp windows. Replication initiation zones are shown as bars.

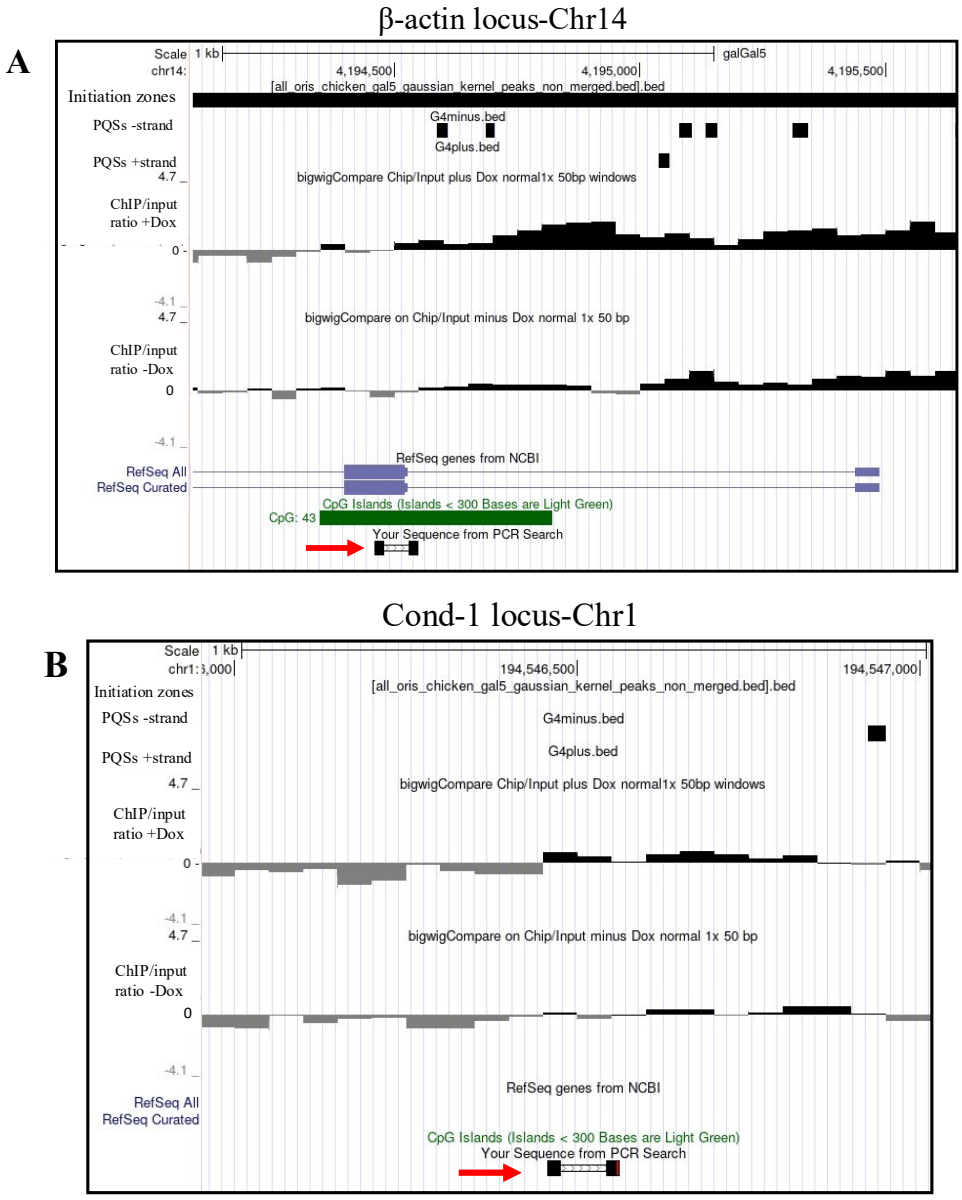

Figure 4.7. Visualization of Chip/Input ratios -/+ Dox in USCS browser for regions analyzed by qPCR. A) β-Actin locus. B) Cond-1 locus. Red arrow indicates position of qPCR amplicon.

4.6 Known G4-binding proteins and replication proteins are enriched in an exploratory G4P-ChIP-MS analysis

Initially, we wanted to use the G4P-APEX2 cell line to perform proximity biotinylation proteomic approach using the APEX2 enzyme. However, the APEX2 control cell line was still under construction (see below). Instead, we decided to perform chromatin immunoprecipitation using the anti-Flag magnetic M2 beads followed by mass spectrometry using label-free quantification (LFQ). We expected that this G4P-ChIP approach would also co-immunoprecipitate proteins with the G4P close to G4s, although we were not sure whether G4P does not compete with G4 binding proteins directly at G4s. No triplicates or duplicates were performed in this first exploratory proteomic analysis as we wanted to test the conditions and validity of this approach. This approach has not been described in the literature to describe a G4-proteome so far, contrary to other approaches (see discussion).

As before crosslinked chromatin was precipitated from -Dox or +Dox treated cells using the same conditions as described for the genomic approaches. Washed IP-M2 magnetic beads (-Dox/+Dox) were sent to the Proteo-Seine Facility at the IJM institute. On beads digestion with trypsin was performed according to the facility's protocols. Peptides for both conditions were subjected to liquid chromatography tandem mass spectrometry (LC/MS/MS) using a timTOF technology, data independent acquisition (DIA) and quantified (false discovery rate <1%). 4333 proteins were detected, 97.7% of them were present in both conditions. The G4P-APEX2 fusion protein was detected with 55 peptides and enriched 24-fold, demonstrating a satisfying bait enrichment with the G4P-ChIP approach. 778 proteins were enriched more than 1.5 and 455 proteins more than 2-fold. Next, I searched for the presence of known G4-binding proteins (G4BP) (Sanchez-Martin, 2023). Out of 32 well characterized G4BP, our G4P-Chip approach found 12 enriched more than at least 1.5-fold), among them Nucleolin (3-fold).

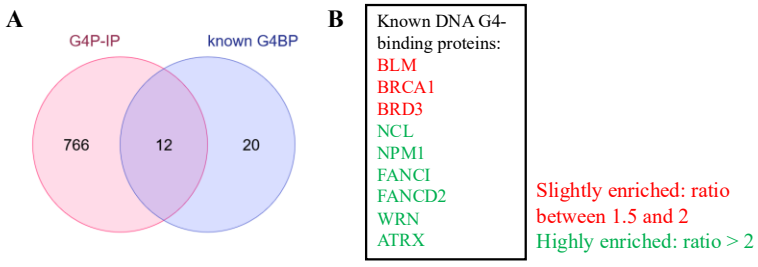

Figure 4.8. G4P-ChIP-MS identifies many known G4 binding proteins. A) Venn diagram showing the overlap between enriched proteins from our G4P-IP and well described G4BPs, listed in a recent review on G4BP from Sanchez et al, 2024. B) List of 12 known G4BP identified in our experiment.

To categorize enriched identified proteins further for their participation in different biological pathways, I performed a pathway and process enrichment analysis for all the enriched proteins (>1.5) (Figure 4.9). I observed that the most significantly enriched pathway is the rRNA processing and general DNA metabolic process, followed by the DNA replication process. Other DNA associated processes identified include DNA repair, transcription, cell cycle, and chromosome organization.

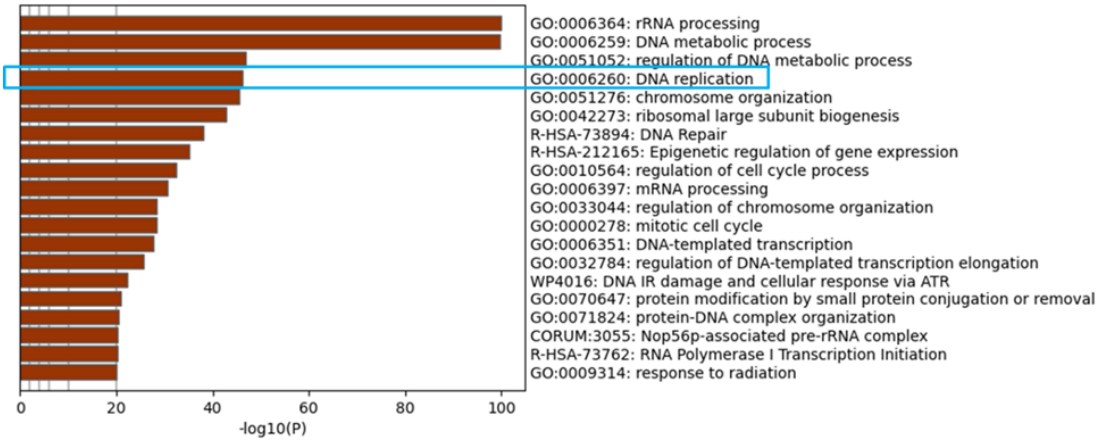

Figure 4.9. Pathway and process enrichment analysis. Top 20 clusters of enriched terms across input of enriched proteins (ratio higher than 1.5) in the +Dox samples, following ontology sources (KEGG Pathway, GO Biological Processes, Reactome Gene Sets, Canonical Pathways, CORUM, and WikiPathways). Log10(P) is the p-value in log base 10. Made by Metascape (metascape.org).

Next, to visualize in more detail enriched proteins in the DNA-associated pathways I manually selected among the over 700 proteins >1.5 fold enriched those involved in DNA replication, repair and recombination pathways, the cell cycle, replication fork stalling, transcription factors, but also proteins involved in chromatin remodeling and epigenetics (Figure 4.10). Interestingly, we find several initiation replication proteins like ORC2, ORC3, GINS, TICRR (Treslin), suggesting a proximity of G4P immunoprecipitated DNA with replication proteins. I observe also transcription factors known to bind at G4 like SP1 and FOXL2, as well as chromatin remodelers and G4 resolving helicases like DDX11, WRN, BLM.

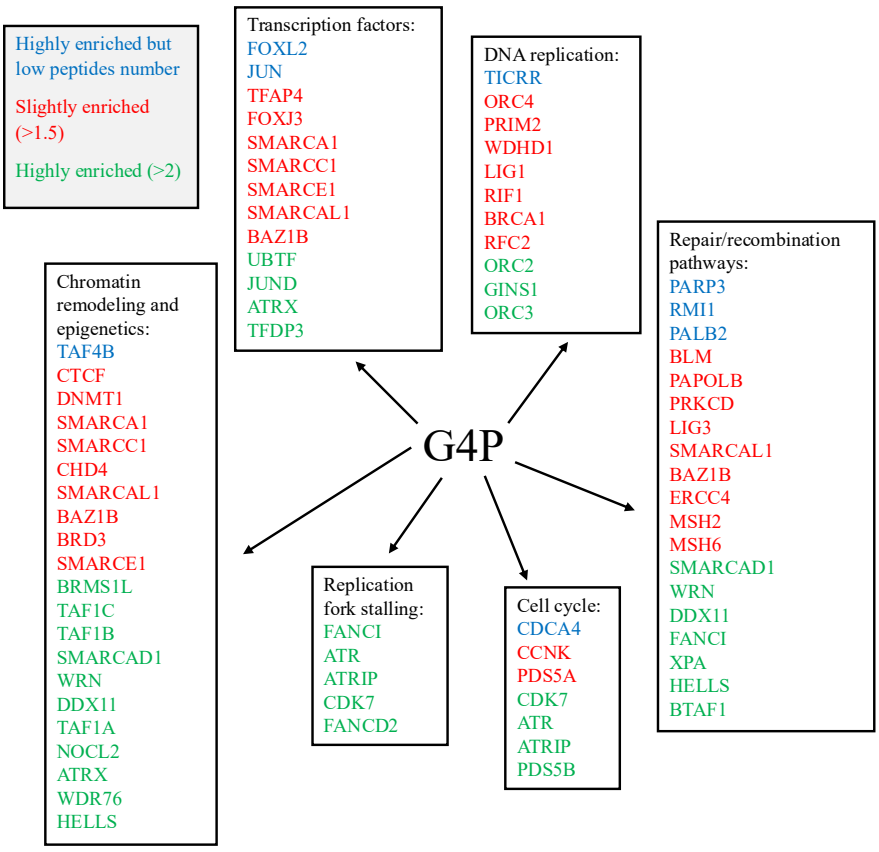

Figure 4.10. G4P-ChIP reveals proximity of G4P with DNA replication proteins and other proteins associated to DNA-pathways. Manually selected enriched proteins by different DNA related processes (blue: proteins highly enriched in the +Dox condition but only present in the amount of 1 or 2 peptides; red: slightly enriched (ratio 1.5-2); green: highly enriched in the +Dox (ratio > 2).

In summary, this first exploratory G4P-ChIP approach seems to confirm several known G4BP and also reveals enrichment of proteins involved in DNA replication and transcription close to G4 recognized by G4P. This is consistent with G4 enrichment in a category of replication origins and in promoters. Therefore, a more quantitative MS analysis using this approach will be done in the future to confirm these preliminary results and to identify new proteins.

4.7 Construction and isolation of APEX2-DT40 cell lines for future proximity biotinylation proteomic approaches

For a proximity biotinylation approach a cell line expressing the APEX2 protein only needed to be constructed as a negative control for future proteomic experiments. The plasmid containing the expression cassette with Flag-tagged APEX2 fused to eGFP under the control of the Dox inducible TRE3G promoter had been constructed before my arrival in the lab. I electroporated the linearized plasmid into WT DT40 cells and selected 8 clones which had stably integrated the cassette under puromycin selection in the genome for further FACS-analysis by eGFP expression after doxycycline addition for 20h at 1 µg/ml. G4P-APEX2 expressing cells of clone 1 were analyzed in direct comparison to select those APEX2-clones with similar an expression level (Figure 4.11). Four clones showed similar eGFP-expression level by FACS.

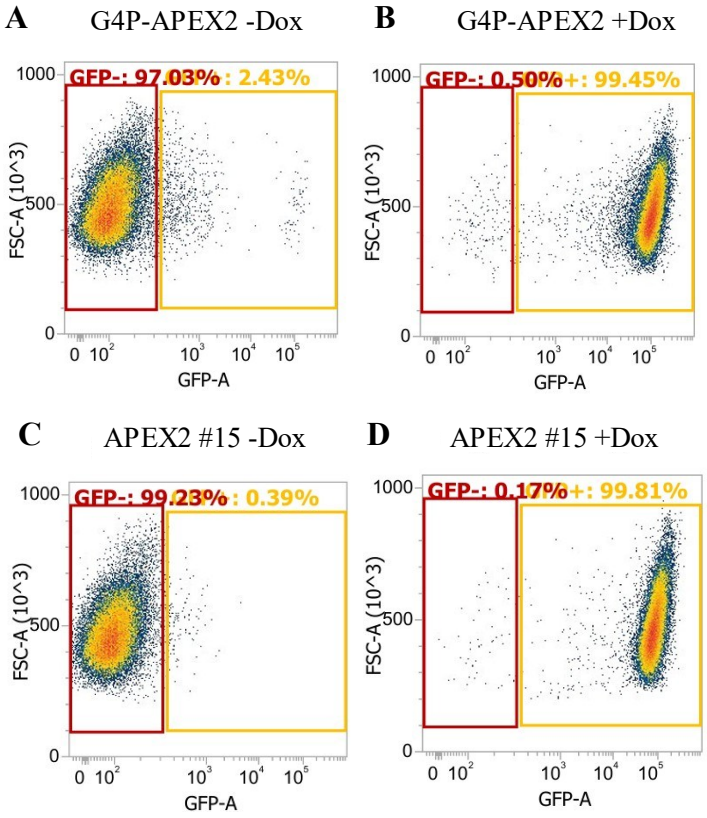

Figure 4.11. Comparison of eGFP expression of FLAG-G4P-APEX2 and a FLAG-APEX2 clone (#15) by FACS analysis. A) FLAG-G4P-APEX2 non-induced clone. B) FLAG-G4P-APEX2 induced clone. C) FLAG-APEX2 non-induced clone 15. D) FLAG-APEX2 induced clone 15.

Next, we analyzed APEX2 expression in these 4 clones by western blot using the anti-flag antibody (Figure 4.12A). As seen before with the induction of G4P-APEX2, the APEX2 was well expressed in all 4 clones upon doxycycline addition. The protein migrated around expected MW at 46 kDa. Tubulin was used as loading control. Flag-tagged APEX2 migrated quicker than the G4P-APEX2 fusion, between 35 and 48 kDa, but a little higher than the predicted MW of 28 kDa. Western blot bands were quantified and normalized by tubulin. (Figure 4.12B). Three APEX2 clones (#2,10,15) showed relatively similar expression levels than the G4P-APEX2 clone. Thus, these clones can serve as correct control cell lines for future proteomic approaches using APEX2.

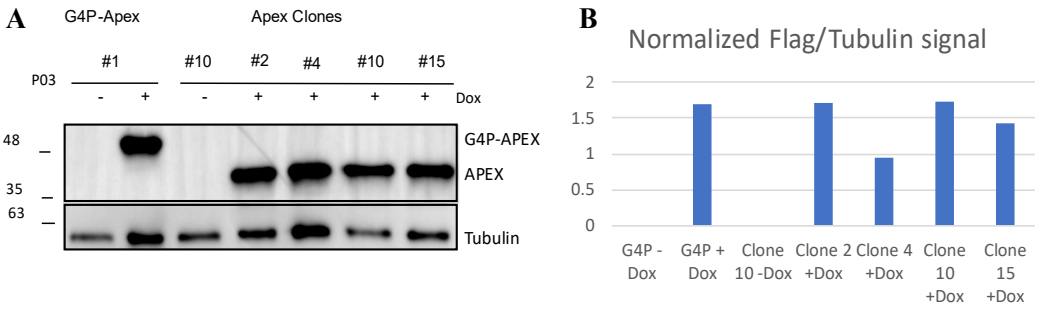

Figure 4.12. APEX2 expression comparison by western blot analysis. A) Cellular extracts of same number of cells of G4P-APEX clone 1 and four different APEX2 clones were prepared in absence and presence of doxycycline and submitted to SDS-PAGE and western blotting, using anti-Flag antibody and anti-tubulin to obtain a loading control for normalization. B) Quantification of G4P-APEX2 and APEX2 bands, normalized by tubulin band intensities. Proteins dimensions are indicated in kDa.

5 Discussion and Conclusions

The regulation of replication origin activation in multicellular organisms is still poorly understood. Strong or efficient replication origins have been associated with structured G4 sequences but how these regulate origin activation and what are the regulating *cis*-factors bound to G4s is unknown. Recently, a new tool was developed to map structured G4s using an artificial G4-probe (G4P) derived from the G4 helicase RHAU expressed in human and chicken cells.

The aim of my internship was to explore the use of this G4-Probe fused to APEX2 to study the proteome around efficient replication origins in DT40 chicken cells. The G4P-APEX2 fusion is intended for proximity biotinylation experiments. However, in my study I used a ChIP-MS approach, since the control APEX2 cell line without the G4P was only ready to use at the end of my internship. The cellular, genomic localization studies and the proteomic approach were therefore carried out with the G4P-APEX2 fusion expressing cell line.

5.1 Expression and nuclear localization of G4P-APEX2 in DT40 cells

I could demonstrate that the G4P fused to APEX2 is expressed upon induction, localized as expected to the nucleus thanks to the nuclear localization sequence added to the G4P-APEX2 in the expression cassette. The fusion protein is found in chromatin fractions and chromatin immunoprecipitation of the flag-tagged G4P-APEX2 using anti-Flag coupled beads could be achieved. This showed that at least a fraction of expressed G4P-APEX2 binds to chromatin. We did not determine whether a fraction of the expressed G4P-APEX2 does not bind to chromatin, but since we overexpress this probe artificially, this could be the case.

Next, I visualized G4P-APEX2 expression by confocal fluorescence microscopy. I could observe two different nuclear patterns. Either G4P-APEX2 immunostained more or less heterogeneously the full area of the nucleus or was restrained to somehow defined large subnuclear structures. This is similar to results obtained by immunolocalization of the G4-probe alone in human cells (Zheng et al., 2020). It would be interesting to study whether these different patterns reflect different cell cycle phases. Furthermore, we could check whether these rather large patches in the nucleus correspond to, for example, the nucleolus by doing co-localization studies using antibodies against known nucleolus-proteins and anti-Flag antibody. G4s can be visualized using a specific DNA G4 antibody, having a strong preference for parallel G4s (Biffi et al., 2013). DNA G4s were detected by the appearance of distinct small nuclear foci. The number of these foci are the highest during S-phase and after addition of G4P stabilization drug. I did not observe any distinct small nuclear foci after G4P-APEX2 overexpression by confocal microscopy. There could be several reasons for

this difference. First, it is possible that the fraction of G4P not bound to chromatin masks the chromatin bound-G4P fraction in immunofluorescence microscopy. Pre-extraction of non-bound G4P-APEX2, prior to fixing, by detergent, and immunostaining might reveal a different staining pattern. Second, the anti-G4 antibody detects only DNA G4s whereas the G4P probably detects both DNA and RNA G4 (see also 5.3), as the RHAU helicase unwinds both RNA and DNA G4s (Heddi et al., 2015).

5.2 Preliminary analysis of G4P-APEX2 enrichment at specific regions in the genome

To analyze where G4P-APEX2 binds genome-wide, I performed a first G4P-ChIP-Seq experiment and visualized ChIP/Input ratios genome wide. Although the signal/noise ratio was rather low, visualization of many regions on different chromosomes showed that G4P peaks are often enriched at CpG islands and co-localized with mapped replication origins. A detailed bioinformatics analysis is now necessary to confirm this preliminary observation and to compare our G4P-APEX peaks with the published G4P peaks from Zheng et al. (Zheng et al., 2020). I also performed a ChIP-qPCR analysis at two loci: one G4 rich versus a control G4 poor region. These two loci had been chosen because the G4-rich β-Actin locus contains PQSs, replication origins (Massip et al., 2019; Poulet-Benedetti et al., 2023) and overlapped also with a G4P peak (Zheng et al., 2020). qPCR primer pairs also normally work well in both regions. However, I only found a relatively low enrichment of G4P-ChIP sequences at the β-actin locus compared to the control locus. This could be due to the fact that the qPCR primers were too far away from G4 rich sequences as shown in Figure 4.7A or to the low ratio signal/background. These first experiments suggest that the ChIP needs to be further optimized for genomic studies, maybe using shorter incubation times with anti-Flag beads or more stringent washing conditions of beads after antibody incubation. Compared to the original study, our ChIP-Seq approach differed in many ways: we used a G4P-APEX2 fusion protein, which could alter G4 binding capacity of the G4P; our G4P expression system, chicken cell line, ChIP protocol, that is adapted to our cell line, are also different.

Many different G4-mapping tools to map G4 genome wide exist, which gave a wide range of G4 numbers formed in human cells, from 10 000 with the G4-ChIP Seq (using anti-G4 antibody (Hänsel-Hertsch et al., 2016)) to 500 000 using G4-Seq (using DNA polymerase stalling (Chambers et al., 2015)) in bulk studies. This illustrates the difficulties of mapping structured G4s *in vivo* which are energetically stable at one hand, but on the other hand are necessarily dynamic and therefore variable because of their threat to genomic stability inside a cell population.

5.3 An exploratory G4P-APEX2-ChIP proteomic showed good enrichment of many known G4-binding proteins

I also performed a first exploratory proteomic approach using G4P-ChIP followed by label-free mass spectrometry using the same protocol for chromatin preparation. Qualitative results show that G4P-APEX2 ChIP leads to an enrichment of 12 known G4 binding proteins from a list of 20 such as nucleolin (NCL), the major nucleolar protein, associated with intra-nucleolar chromatin and pre-ribosomal particles. Nucleolin plays a role in pre-rRNA transcription and ribosome assembly. I also found enriched nucleophosmin (NPM1), a nucleolar phosphoprotein and many other proteins involved in ribosome biogenesis which was the most abundant pathway in the pathway enrichment analysis. We ignore if this explains our localized nuclear staining pattern for the G4P in the immunofluorescent studies in some cells, but it would be interesting to investigate. We observe as well an enrichment of many G4 helicases such as the Werner helicase (WRN), the Bloom-Helicase (BLM) and the helicase DEAH-box DDX11, that has been implicated recently to resolve G4s during DNA synthesis by interaction with fork protection proteins Timeless and Claspin (Lerner et al., 2020) to ensure DNA synthesis across G4s. Other known G4 helicases, we did expect, like Pif1 or FANCJ, have not been found enriched.

Most interestingly for us, we found some enriched replication initiation proteins, such as ORC2, ORC4, Treslin, and GINS. ORC1-6 is the initiator and is composed of a complex of 6 subunits. ORC1 binds preferentially to G-quadruplex (G4)-preferable G-rich RNA or single-stranded DNA (ssDNA) (Hoshina et al., 2013). TICRR, also called Treslin, together with its partner MTBP are initiation factors necessary for the loading of GINS and CDC45 on the MCM complex to constitute the replicative helicase. MTBP binds to G4s *in vitro* and MTBP binding sites contain PQSs (Kumagai and Dunphy, 2020). Finally, we also found Rif1 enriched, which has been shown to bind to G4 motifs *in vitro* (Kanoh et al., 2015). Rif1 is a multi-functional protein, implicated in DNA repair and recombination in multicellular organisms. During DNA replication, it is known to target the phosphatase PP1 to origins which inhibits the activation of the replicative helicase and preferentially inhibits activation of late firing origins. In yeast, Rif1 binding sites overlap with sites of dormant origins, which are origins inhibited during normal S-phase, but which can be activated upon replication stress (Kanoh et al., 2015). Rif1 binding sites in mice cells are found also overlapping with TSS, CGIs and a subset of early firing origins. Altogether, my results from this first proteomic study are interesting as we found both known G4BP and replication initiation proteins enriched. This suggests that the fusion of G4P to APEX2 still allows G4 binding of the fusion protein. To investigate whether the G4 binding is altered in the G4P-APEX2 fusion protein, a direct comparison with a proteomic approach of the G4P alone would be interesting.

Recently, several studies have used different approaches to identify the proteome around G4s. Therefore, I compared our preliminary results with two other studies (Figure 5.1). The group led by Lu (Lu et al., 2024) developed an approach using miniTurbo, a biotin ligase, fused to one RHAU23 peptide to generate a new bifunctional probe, G4PID and identify a G4 proteome using whole human cellular extracts. This approach is similar to ours, but I prepared chromatin for IP, not extracts from whole cells. I found that 7 enriched proteins were in common between this study and my results. The proteins mainly belong to the ribosome biogenesis pathway, but no nucleolin was found enriched in their study. A second study (Zhang et al., 2021) used a cell permeable G4-ligand, a pyridostatin derivative, tethered to a photo-activable group and an alkyne handle. Ultraviolet irradiation triggers the proximity capture of co-binding G4 interacting proteins in human cells. Using this technique, they analyzed nuclear proteins only. I found 20 proteins overlapping between this study and my results, among these are nucleolin, nucleophosmin, many helicases and proteins of the ribosome biogenesis pathway. The overlap between the two published studies themselves was low (4) despite the fact that they used the same human cell line (HEK293T). The only protein common to all three studies is GAR1, H/ACA ribonucleoprotein complex subunit 1, which is part of the small nucleolar ribonucleoproteins family. These little overlaps between G4 proteomic studies probably reflect the different approaches and cellular materials (cellular extracts versus nuclear extracts).

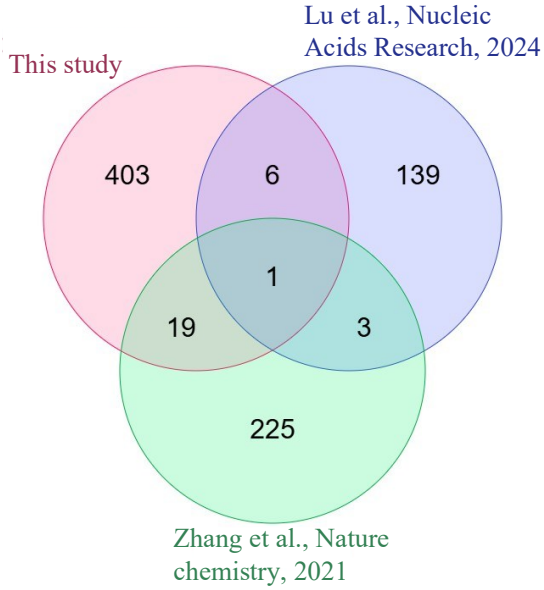

Figure 5.1. Comparison of our G4P-proteome with two other G4 proteomic studies. Venn-Diagram depicts proteins found in my results using G4P-APEX2-ChIP from chicken DT40 cells (proteins enriched >1.5) with proximity biotinylation using RHAU23-miniTurbo in human cells (HEK293T) (Lu et al., 2024) and photoactivable chemical-co-binding and cross-linking in human cells (HEK293T) (Zhang et al., 2021).

In conclusion, the results of the proteomic study using G4P expressed in DT40 cells indicate that G4P can be used as a tool to explore proteins at, or around PQSs, and at PQS-cluster containing replication origins, since I found known G4BP and also replication initiation proteins enriched in my G4P ChIP- MS approach. We have to confirm these very preliminary results with a quantitative experiment with triplicates samples for statistical robustness. This G4P-ChIP MS approach could then also allow the identification of novel players regulating replication origins around clustered PQSs. We would also need to construct a cell line expressing only flag-tagged G4P to compare to the G4P-APEX2 fusion. This could be used as a second complementary approach to the proximity biotinylation approach with APEX2, which we initially planned to use and which we are now ready to implement with the construction of the APEX2 control cell line. The genome-wide G4 mapping approach in DT40 cells using G4P shows some enrichment at CGIs but a more detailed bioinformatic analysis is necessary. This approach will also need further experimental optimization due to a relatively low signal/background ratio.

It would be also interesting to investigate the G4 proteome at different stages of the cell cycle (G1, S, G2) to decrease cell heterogeneity inside the cell population using either synchronization or elutriation protocols.

We were intrigued to see that the results from my preliminary proteomic study are closer to what we expected and also more convincing than the results from the genomic study using the same ChIP protocol for G4P-IP. Single-cell studies revealed genome-wide only around 700 G4 peaks (Hui et al., 2021), compared to 20 000 peaks using the same CUT&Tag mapping technique (Li et al., 2021) in a cell population approach. This suggests that there is a large cell-to-cell variability in G4 positioning, and also that only a small fraction of all PQSs folds into a G4 structure. Maybe the G4 proteome is somehow more robust to analyze than the G4 genomic binding sites inside a given cell population and could therefore further contribute to our understanding of the G4 function in cells.

6 Bibliography

Besnard, E., Babled, A., Lapasset, L., Milhavet, O., Parrinello, H., Dantec, C., Marin, J.-M., Lemaitre, J.-M., 2012. Unraveling cell type–specific and reprogrammable human replication origin signatures associated with G-quadruplex consensus motifs. Nat. Struct. Mol. Biol. 19, 837–844. https://doi.org/10.1038/nsmb.2339

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

Brázda, V., Hároníková, L., Liao, J.C.C., Fridrichová, H., Jagelská, E.B., 2016. Strong preference of BRCA1 protein to topologically constrained non-B DNA structures. BMC Mol. Biol. 17, 14. https://doi.org/10.1186/s12867-016-0068-6

Cadoret, J.-C., Meisch, F., Hassan-Zadeh, V., Luyten, I., Guillet, C., Duret, L., Quesneville, H., Prioleau, M.-N., 2008. Genome-wide studies highlight indirect links between human replication origins and gene regulation. Proc. Natl. Acad. Sci. 105, 15837–15842. https://doi.org/10.1073/pnas.0805208105

Cayrou, C., Coulombe, P., Puy, A., Rialle, S., Kaplan, N., Segal, E., Méchali, M., 2012. New insights into replication origin characteristics in metazoans. Cell Cycle 11, 658–667. https://doi.org/10.4161/cc.11.4.19097

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

Chen, M.C., Tippana, R., Demeshkina, N.A., Murat, P., Balasubramanian, S., Myong, S., Ferré-D'Amaré, A.R., 2018. Structural basis of G-quadruplex unfolding by the DEAH/RHA helicase DHX36. Nature 558, 465–469. https://doi.org/10.1038/s41586-018-0209-9

Fenouil, R., Cauchy, P., Koch, F., Descostes, N., Cabeza, J.Z., Innocenti, C., Ferrier, P., Spicuglia, S., Gut, M., Gut, I., Andrau, J.-C., 2012. CpG islands and GC content dictate nucleosome depletion in a transcription-independent manner at mammalian promoters. Genome Res. 22, 2399–2408. https://doi.org/10.1101/gr.138776.112

Guilbaud, G., Murat, P., Wilkes, H.S., Lerner, L.K., Sale, J.E., Krude, T., 2022. Determination of human DNA replication origin position and efficiency reveals principles of initiation zone organisation. Nucleic Acids Res. 50, 7436–7450. https://doi.org/10.1093/nar/gkac555

Hänsel-Hertsch, R., Beraldi, D., Lensing, S.V., Marsico, G., Zyner, K., Parry, A., Di Antonio, M., Pike, J., Kimura, H., Narita, M., Tannahill, D., Balasubramanian,

S., 2016. G-quadruplex structures mark human regulatory chromatin. Nat. Genet. 48, 1267–1272. https://doi.org/10.1038/ng.3662

Heddi, B., Cheong, V.V., Martadinata, H., Phan, A.T., 2015. Insights into G-quadruplex specific recognition by the DEAH-box helicase RHAU: Solution structure of a peptide–quadruplex complex. Proc. Natl. Acad. Sci. 112, 9608–9613. https://doi.org/10.1073/pnas.1422605112

Hoshina, S., Yura, K., Teranishi, H., Kiyasu, N., Tominaga, A., Kadoma, H., Nakatsuka, A., Kunichika, T., Obuse, C., Waga, S., 2013. Human Origin Recognition Complex Binds Preferentially to G-quadruplex-preferable RNA and Single-stranded DNA. J. Biol. Chem. 288, 30161–30171. https://doi.org/10.1074/jbc.M113.492504

Hui, W.W.I., Simeone, A., Zyner, K.G., Tannahill, D., Balasubramanian, S., 2021. Single-cell mapping of DNA G-quadruplex structures in human cancer cells. Sci. Rep. 11, 23641. https://doi.org/10.1038/s41598-021-02943-3

Jana, J., Mohr, S., Vianney, Y.M., Weisz, K., 2021. Structural motifs and intramolecular interactions in non-canonical G-quadruplexes. RSC Chem. Biol. 2, 338–353. https://doi.org/10.1039/D0CB00211A

Kanoh, Y., Matsumoto, S., Fukatsu, R., Kakusho, N., Kono, N., Renard-Guillet, C., Masuda, K., Iida, K., Nagasawa, K., Shirahige, K., Masai, H., 2015. Rif1 binds to G quadruplexes and suppresses replication over long distances. Nat. Struct. Mol. Biol. 22, 889–897. https://doi.org/10.1038/nsmb.3102

Kobayashi, S., Fukatsu, R., Kanoh, Y., Kakusho, N., Matsumoto, S., Chaen, S., Masai, H., 2019. Both a Unique Motif at the C Terminus and an N-Terminal HEAT Repeat Contribute to G-Quadruplex Binding and Origin Regulation by the Rif1 Protein. Mol. Cell. Biol. 39, e00364-18. https://doi.org/10.1128/MCB.00364-18

Kumagai, A., Dunphy, W.G., 2020. Binding of the Treslin-MTBP Complex to Specific Regions of the Human Genome Promotes the Initiation of DNA Replication. Cell Rep. 32, 108178. https://doi.org/10.1016/j.celrep.2020.108178

Lam, S.S., Martell, J.D., Kamer, K.J., Deerinck, T.J., Ellisman, M.H., Mootha, V.K., Ting, A.Y., 2015. Directed evolution of APEX2 for electron microscopy and proximity labeling. Nat. Methods 12, 51–54. https://doi.org/10.1038/nmeth.3179

Lerner, L.K., Holzer, S., Kilkenny, M.L., Šviković, S., Murat, P., Schiavone, D., Eldridge, C.B., Bittleston, A., Maman, J.D., Branzei, D., Stott, K., Pellegrini, L., Sale, J.E., 2020. Timeless couples G-quadruplex detection with processing by DDX 11 helicase during DNA replication. EMBO J. 39, e104185. https://doi.org/10.15252/embj.2019104185

Li, C., Wang, H., Yin, Z., Fang, P., Xiao, R., Xiang, Y., Wang, W., Li, Q., Huang, B., Huang, J., Liang, K., 2021. Ligand-induced native G-quadruplex stabilization impairs transcription initiation. Genome Res. 31, 1546–1560. https://doi.org/10.1101/gr.275431.121

Lu, Z., Xie, S., Su, H., Han, S., Huang, H., Zhou, X., 2024. Identification of G-quadruplex-interacting proteins in living cells using an artificial G4-targeting biotin ligase. Nucleic Acids Res. 52, e37–e37. https://doi.org/10.1093/nar/gkae126

Massip, F., Laurent, M., Brossas, C., Fernández-Justel, J.M., Gómez, M., Prioleau, M.-N., Duret, L., Picard, F., 2019. Evolution of replication origins in vertebrate genomes: rapid turnover despite selective constraints. Nucleic Acids Res. 47, 5114–5125. https://doi.org/10.1093/nar/gkz182

Mukundan, V.T., Phan, A.T., 2013. Bulges in G-Quadruplexes: Broadening the Definition of G-Quadruplex-Forming Sequences. J. Am. Chem. Soc. 135, 5017–5028. https://doi.org/10.1021/ja310251r

Nasheuer, H.P., Meaney, A.M., 2024. Starting DNA Synthesis: Initiation Processes during the Replication of Chromosomal DNA in Humans. Genes 15, 360. https://doi.org/10.3390/genes15030360

Picard, F., Cadoret, J.-C., Audit, B., Arneodo, A., Alberti, A., Battail, C., Duret, L., Prioleau, M.-N., 2014. The Spatiotemporal Program of DNA Replication Is Associated with Specific Combinations of Chromatin Marks in Human Cells. PLoS Genet. 10, e1004282. https://doi.org/10.1371/journal.pgen.1004282

Poulet-Benedetti, J., Tonnerre-Doncarli, C., Valton, A.-L., Laurent, M., Gérard, M., Barinova, N., Parisis, N., Massip, F., Picard, F., Prioleau, M.-N., 2023. Dimeric G-quadruplex motifs-induced NFRs determine strong replication origins in vertebrates. Nat. Commun. 14, 4843. https://doi.org/10.1038/s41467-023-40441-4

Prioleau, M.-N., MacAlpine, D.M., 2016. DNA replication origins—where do we begin? Genes Dev. 30, 1683–1697. https://doi.org/10.1101/gad.285114.116

Rhee, H.-W., Zou, P., Udeshi, N.D., Martell, J.D., Mootha, V.K., Carr, S.A., Ting, A.Y., 2013. Proteomic Mapping of Mitochondria in Living Cells via Spatially Restricted Enzymatic Tagging. Science 339, 1328–1331. https://doi.org/10.1126/science.1230593

Samaniego-Castruita, D., Han, I., Morgan, R.C., Carpenter, S., Williams, B., Chakraborty, A., Radhakrishnan, I., Ay, F., Myers, S.A., Rao, A., Shukla, V., 2025. CTCF directly binds G-quadruplex structures to regulate genome topology and gene expression. https://doi.org/10.1101/2025.02.03.636329

Sanchez-Martin, V., 2023. DNA G-Quadruplex-Binding Proteins: An Updated Overview. DNA 3, 1–12. https://doi.org/10.3390/dna3010001

Schiavone, D., Guilbaud, G., Murat, P., Papadopoulou, C., Sarkies, P., Prioleau, M., Balasubramanian, S., Sale, J.E., 2014. Determinants of G quadruplex-induced epigenetic instability in REV 1-deficient cells. EMBO J. 33, 2507–2520. https://doi.org/10.15252/embj.201488398

Schult, P., Paeschke, K., 2021. The DEAH helicase *DHX36* and its role in G-quadruplex-dependent processes. Biol. Chem. 402, 581–591. https://doi.org/10.1515/hsz-2020-0292

Song, D., Luo, J., Duan, X., Jin, F., Lu, Y.-J., 2025. Identification of G-quadruplex nucleic acid structures by high-throughput sequencing: A review. Int. J. Biol. Macromol. 297, 139896. https://doi.org/10.1016/j.ijbiomac.2025.139896

Valton, A.-L., Prioleau, M.-N., 2016. G-Quadruplexes in DNA Replication: A Problem or a Necessity? Trends Genet. 32, 697–706. https://doi.org/10.1016/j.tig.2016.09.004

Zhang, X., Spiegel, J., Martínez Cuesta, S., Adhikari, S., Balasubramanian, S., 2021. Chemical profiling of DNA G-quadruplex-interacting proteins in live cells. Nat. Chem. 13, 626–633. https://doi.org/10.1038/s41557-021-00736-9

Zheng, K., Zhang, J., He, Y., Gong, J., Wen, C., Chen, J., Hao, Y., Zhao, Y., Tan, Z., 2020. Detection of genomic G-quadruplexes in living cells using a small artificial protein. Nucleic Acids Res. 48, 11706–11720. https://doi.org/10.1093/nar/gkaa841