Deciphering the Evolutionary Conservation and Functional Role of G4-Binding Proteins in Genome Stability and Cancer Progression

Mehul Rathi

Received November 28, 2024
Accepted May 29, 2025
Electronic access July 15, 2025

This work further explores the part of identified G-quadruplex (G4) binding proteins; Nucleolin, HNRNPA1, DHX36, BRCA1, FANCJ, and WRN in processes ranging from replication, repair, transcription, to the impact of cellular stresses. Thus, this research uses sequence alignment and Missense Tolerance Ratio (MTR) in humans, mice, chickens, and zebrafish to expose evolutionary conservation and functional similarity of these proteins. The alignments reveal information regarding protein region conservation, helping to identify which amino acids could be important for the proteins' function, and which could be involved in G4 binding. The MTR is used to recognize that there are some regions in such proteins that are very conserved and relatively intolerant to missense mutation, which highlights protein regions that could be essential for its functions. Based on these findings, we propose that the relationships of these proteins with G4 structures could be important in regulating cellular activities and could be involved in tumor advancement. The study advances the knowledge of molecular targets in regulating the cellular reaction to disruptions in DNA integrity and provides drug targets for tumor therapy. Therefore, it does contribute to the development of oncology since it illuminates the key factors causing cancer progression and begins to expound additional approaches to treatment.

Introduction

G4s are non-canonical secondary nucleic acid structures formed by Guanine-rich sequences whereby guanine tetrads stack into stable four stranded structures. G4s have been studied for the past two decades with a focus on their role in acting as molecular switches to regulate some of the essential cellular processes such as DNA replication, gene transcription, and cell aging[1]. Notably, G4s are also under a great focus of research with regards to their probable involvement in the enhancement of tumor growth and progression thus considering them targets for new therapies. Despite the established participation of G4s in disease progression, conclusive proof of the structures' direct contribution to cancer onset is still lacking[2]. This leads us to the central focus of our research: What is the role of G4's in DNA replication and cancer cell proliferation?

Therefore, it is necessary to study the proteins that interact with G4 structures to understand the biological importance of them. These G4 binding proteins act as facilitators or regulators of G4 structures through promoting resolution or stabilization. For example, Nucleolin and HNRNPA1 are both crucial proteins in the regulation of transcription and processing of RNA and others, like BRCA1 and WRN, are important for genome integrity by DNA repair. Searching their conservation across species not only helps to identify universally essential domain but also to learn something about evolutionary pressures that have shaped their function. Computational tools are employed on this study to examine sequence similarity and functional constraints using MTR scores to find the positions of missense intolerance, potentially correlated with the binding or catalytic domains.

The formation of G4s in DNA, for example, might create natural chokepoints for DNA replication forks, leading to replication stress – a common state of cells in cancers[2]. During transcription, G4s can modulate patterns of gene expression, and can affect expression levels of oncogenes and tumor suppressor genes, leading to changes in the rate of cancer cell growth[3]. The intricate structure suggests its involvement in cellular aging through affecting the telomeres' stability and functionality. This in turn helps, most likely, to oncogenesis and cellular senescence. Cellular senescence is a process in which a cell ages and stops dividing permanently, but does not die[4].

Given their pervasive effects on genomic stability and cellular metabolism, G4s are increasingly being recognized as suitable drug targets, with molecules designed to stabilize or destabilize G4 structures showing potential for controlling cancer cell growth, among others, providing a novel avenue in the treatment of cancer[5].

Biological Significance of G4s

G4s, with their unique four-stranded guanine-rich structures, are highly enriched in regulatory parts of the genome, including at the ends of chromosomes (telomeres) and at the promoter regions of oncogenes[1]. G4s are non-randomly located in the genome, suggesting that they play an important role in both the

structural architecture and the transcription regulation of the genome[2].

G4s are functional by changing the landscape of the genome, stabilizing and destabilizing specific genome regions under physiological conditions[2]. In the case of telomeres, G4 structures are protective formations, which weaken the chromosomal ends and the genomic instability that predisposes to cellular senescence and oncogenic transformation[3]. This protective mechanism is of great importance to sustain the integrity of telomeric structures, particularly in cells that divide rapidly, where telomere stability is necessary to maintain their replicative potential[6].

G4s are also found in gene promoters, and they can affect the access of transcription factors and other regulatory proteins to their binding sites. Stabilizing or destabilizing transcription factor binding at gene promoters can either increase or impair gene expression, directly affecting cellular growth and differentiation pathways[7]. These regulatory functions make G4s a relevant part of cellular machinery that directly impacts growth control and response to developmental cues[8].

Yet, the stability brought by G4s is offset by its ability to promote genomic instability. Stable G4s can obstruct DNA replication by creating physical barriers that allow DNA replication forks to become wrapped, forming a physical blockade, like a traffic jam for the machinery of replication, ultimately inducing replication stress – a type of DNA damage that can lead to genomic instability, a hallmark of many cancers[9,10]. Moreover, fluctuating formation and dissociation of G4s can trigger transcriptional dysregulation, which can cause a downstream disruption in cellular functions[11].

This dual, facilitative-obstructive character further underscores the pivotal nature of G4s in sustaining a dynamic cellular homeostasis, with their functions being an integral part of the cellular stress response in reaction to both intracellular and extracellular disruptions – as transcription and genomic integrity modulators. Through these roles, G4s are fundamental to maintaining a dynamic equilibrium – essentially, cellular resilience – through enabling the cell to deal with environmental and physiological stresses, ensuring homeostasis and thus cellular health[1].

G4s and DNA Replication

These structures occur when DNA adopts non-B (Non-B DNA) forms, aside from the classic right-handed double helix known as B-DNA. Among the non-B DNA structures, there exists one called G4. This structure is made from the DNA strand part containing a large amount of guanine. These guanine bases can lie in a specific square fashionable, which helps in forming a four-stranded structure. This arrangement is most favorable and very stable. Stable structures such as this have the potential to impact genomic processes such as replication and transcription[1].

During DNA replication, G4 structures are inherently more resistant to the unwinding activities performed by the replication machinery, they create physical and therefore kinetic barriers for the replication fork, stalling the replication enzymes and inhibiting their progress. Stalled replication forks are dangerous as they can lead to replication stress. What is replication stress? It occurs when replication fork progression slows down or stops. If not repaired, it leads to double-strand breaks (DSB's), which are the most dangerous lesions, resulting in genomic instability, which is also a hallmark of cancer and is associated with increased mutation rates and chromosomal alterations[2].

These disruptions are managed by specialized helicases responsible for unwinding G4 structures, allowing DNA replication to progress as normal. Successful resolution of the G4 allows replication to take place without interruption. This seems like an important part of the genomic repair kit, managing these structures that are generally inhibitory to replication. However, since G4s are easily unwound and resolves by specific helicases, it points out that, effective though they are in preventing DNA replication, their regulation and quick resolution are, equally, necessary. This controlled management is necessary for the homeostasis of the cells and also for replication to occur without disturbing the genome[2].

Beyond causing problems, G4s have also been shown to act as regulatory features of the genome, influencing gene expression. This regulation is achieved by altering the binding sites where specific transcription factors and other regulatory proteins attach to G4s in chromosomal DNA. These changes in binding sites affect the transcriptional activity of crucial genes, which are essential for directing various developmental processes and cellular responses to environmental stimuli. These examples highlight how G4s perform important roles in genetic control of the cell[12].

Furthermore, G4s might help to establish replication origins in metazoans. According to this working hypothesis, G4s can help to 'organize' or 'stabilize' the structures and components needed for the initiation of DNA replication, thus facilitating the replication fork initiation – a key step of DNA replication and cell doubling[13].

G4s also contribute to chromatin architecture in another way. They promote topologically associating domains (TADs) and topologically associating genomic domains (TGDs) that are necessary for proper genomic structural organization. Long-range genomic interactions induced by G4s allow us to understand how structural genomic features contribute to something as critical as genomic stability and gene regulation, which in turn impacts every aspect of cellular function[14].

This highlights their multifunctionality, which in turn also underlines their crucial role in genomic homeostasis and regulation; G4s ensure the integrity of chromosomal ends through their replication-challenging functions, while their regulatory and structural functions provide a means of avoiding or reversing viral infections and pathological or detrimental processes. Such multifunctionality is also a characteristic of other pro-

teins that have been envisioned as potential targets for cancer therapy; for example, Lamin which is a structural protein, bromodomain which is a core component of transcription regulation and elongation complexes, ankyrin which is a modulator of protein interactions and many others. The possibility of modulating the functions of G4s and exploiting their intrinsic properties and cellular/organismal functions opens new avenues for cancer treatment [3].

G4s and Gene Expression Modulation

G4s, which consist of four-stranded nucleic acid structures with a guanine-rich backbone, have been shown to affect gene expression by regulating accessibility and activity at DNA regulatory regions. They are not randomly scattered throughout the genome. Rather, G4s are enriched within gene-regulatory regions that govern the expression of genes involved in cancer, and importantly focus on the control of key oncogenes [1].

Ideally, a G4-specific antibody, that is designed to bind only G4 structures (for example, PB22) should be tethered to their G4 signature. This antibody allowed the enrichment of genomic DNA fragments containing the G4s in question, which were then sequenced – providing a sort of roadmap of G4 locations across the genome. This type of mapping not only proved the existence of the G4s but allowed the discovery of unprecedented resolution of their locations and the delineation of potential function in situ. This approach highlights that having the right tools is crucial for identifying and understanding the loci of interest (LOIs) within complex DNA structures, which play a significant role in regulating genomes [15].

Sequencing results showed that G4s are found frequently next to transcription start sites upstream of genes, where they're likely to bind to transcription factors and other regulatory proteins in the cell. If the cell wants to modulate the rate of transcription of a gene – thereby turning genes on or off – forming or disassembling these structures would be a rapid way to achieve that. It is important that genes' activity and inactivity are controlled well in cancer since the development of cancer is characterized by irregular activation of certain genes and inactivation of others which promote rapid cell division. [5].

The findings had implications beyond a simple biological insight: they indicated that G4s contributed to the transcriptional regulation of the cancer genes investigated. By altering gene expression, the DNA G4s were beginning to modify the cellular pathways that control cell division, apoptosis, and metastasis. They could therefore be targets for therapeutic intervention in terms of modulating their formation or stability, providing new ways to control the progress of a malignancy [16].

With this work, a new landscape for drug discovery for G4 targeting therapies has been established. Exquisite molecules can be designed to bind and modulate G4 in cancer genes specifically and selectively inhibit cancer in a cell-specific fashion. In this way, it might be possible to enhance the potency of treatments and reduce toxicity to normal cells. In the future, we may have more precise ways to fight the battle with cancer [1].

Protein Interactions with G4s

As important as proteins are, visualizing how they engage with G4 structures is vital to explain the detailed roles that these special nucleic acid conformations play in cellular processes. G4's are unique, four-stranded structures in DNA or RNA that are formed by sequences rich in guanine bases. Opposite guanines in a coordinated stack can create connections known as Hoogsteen hydrogen bonds. In these bonds, a hydrogen ion traverses space to link two complementary molecules in an atypical fashion [1]. These structures have emerged as important players not just in the genome but for cellular functions such as DNA replication, repair, transcription and translation – because they engage with proteins that drive these processes. This fact underscores the importance of G4s as dynamic and integral players in genomic stability and gene expression [16].

This interaction also highlights the significance of G4's as dynamic and integral players in genomic stability and gene expression. The word 'dynamic' stresses their active and adaptable involvement in the regulation of DNA replication, as G4's can either hinder the replication machinery or expedite replication, depending on their resolution or stabilization by specific helicases [16]. 'Integral' emphasizes their role as core players in such processes as gene expression, specifically because G4's formation in promoter regions can affect transcription regulation either through upregulation or downregulation of transcriptional activity [1].

In binding with the transcription factors and controlling the formation of the transcriptional complex, G4's determines the transcriptional programs of cells, which in turn determine the gene expression patterns vital for cell organization and an appropriate response to stimuli. Additionally, G4's ensures stability and protection of genomic regions, especially of telomeres and gene regulation sites, hence preserving a healthy cell from genetic instabilities and mutations as well as chromosomal abnormalities [17].

This study reflects upon the evolutionary conservation shown here, and thus suggests that G4 binding proteins are critical for genomic stability. When it comes to proteins like HNRNPA1 and DHX36, they are highly conserved across mammals, as they play a crucial role in resolving mammalian G4s responsible for transcription and replication. On the other side, though functionally important, missense tolerance of proteins such as BRCA1 and WRN is more diverged and variable, and their roles might be subjected to species specific regulation or adaptation. This knowledge of the extent of conservation in the interaction domains of these proteins helps to build better mechanistic models and also in their propensity to being therapeutic targets.

For example, WRN helicase conserved low MTR regions within ATPase domain (the ATPase domain unwinds G4s and is used in repairing damaged DNA), represent evolutionary constraints of this G4 unwinding function. It is critical information to design small molecule inhibitors or activators that bind these specific protein G4 interaction sites.

Other proteins that bind to G4 structures can have a profound effect on how such structures are stabilized or unwound within a cell either promoting or preventing/aggravating the processes linked with such formations. Thus, depending on the type of protein interacting with G4 structures, genetic processes related to transcription and replication can be modulated. For instance, Nucleolin and Heterogeneous Nuclear Ribonucleoprotein A1 (HNRNPA1) have been reported to interact with G4's thus changing their stability affecting the cells response to replication stress [14].

During replication of DNA G4 structures present within the helix pose a menace to the replication machinery because they can hamper replication forks. One of the protein functions includes preventing these structures from unwinding, an aspect that could potentially lessen replication stress because the replication machinery will not come across more complex G4s. On the other hand, proteins that are apt to unfold the G4 structures like specific helicase are instrumental in the timely unwinding of such formations hence the continuity of the replication fork and efficiency in DNA replication [16].

When speaking of transcription, it is important to mention that binding of G4 structures with proteins plays a crucial role in regulation of gene expression. Some of the proteins can either activate or repress the transcription process by altering the ability of transcription factors and RNA polymerase to interact with the DNA template. This modulation is essential for the operation of a cell in signaling and the general management of physiological processes. For instance, Nucleolin can bind to G4 structures located in promoter sequences of specific oncogenes which might result in changes in the transcription and consequently affect the cancer process and differentiation of cells [15].

Here, the focus is on the interactions between proteins and G4s where it goes beyond the cellular functions to disease conditions, including cancer. Through impacts on DNA replication and transcription, these events are central to cell growth and sustenance. It was found that aberrant regulation of protein-G4 interactions can cause genomic instability that is the characteristic feature of cancer cells; therefore, the interactions could be potential targets for the therapies.

Just as proteins are involved in the formation and stability of G4's, it shows that the cell possesses a very fine control over its genetic material. These engagements are not just relevant for the running of day-to-day cellular functions but also forms the basis for developing a framework for understanding diseases involving cell division and mutation for instance carcinoma. Further study of these events is essential for the chances in creating new effective therapies that would be based on the interaction of G4's and proteins and could lead to the creation of new effective treatments for cancers and other genetic illnesses [1].

G4's are commonly observed at somatic break points; these are regions which are believed to be vulnerable to break and rearrangement as the cell divides. These regions are very important in cancer because mutations that occur in the regions may either activate oncogenes or inactivate tumor suppressor genes. It has been found out by scientific studies that G4 structures for the physical characteristics and stability put mechanical pressure on the DNA strand during the replication and transcriptase processes. This stress can hinder the action of replication forks, increase the formation of DNA double strand breaks and raise the tendency to mutations [1].

There are numerous reports on the interconnection between replication stress induced by G4, and cancer advancement. Some investigations show that, when replication forks meet with untreated G4 structures, it results in replication fork stalling and collapses which constitutes a major source of genomic instability and is the driving force of cancer. Specific proteins that can dissolve G4s are essential to alleviate this stress and they include helicases. However, in the cancer cells, the expression or function of these helicases becomes faultier and contributes to the enhancement of this issue and the introduction of instability to the cancer cells' genomes [16].

G4s which were previously identified in replication now have functions in the transcription process also by altering the conformation of DNA. This modulation impacts on the ability of transcription factors as well as other regulatory proteins to bind DNA. For instance, the research work has it that G4 structures can either promote or suppress the transcription of some genes basing on its position regarding gene promoters. This duality is in accordance with the fact that members of the G4 family can act simultaneously as 'gene activator' and 'gene repressor', which may play a major role in cancer, as the regulation of gene expression is often disturbed in this disease [1].

Since G4s are important in genomic stability and gene expression, they are considered as potential anticancer targets. The molecules that can selectively interact with G4 structures can either stabilize, or destabilize these formations, thus targeting the altered gene expression patterns characteristic of cancer cells. Based on this approach, many research papers have been written where G4-stabilizing or G4-destabilizing agents have proved to hold potential for directly affecting cancer cells growth and survival and thus providing novel therapeutic targets in cancer [14].

Therefore, further studies on G4's will be vital as the researchers work on the improvement of cancer therapies. More research should be carried out to determine the conditions under which different G4s promote cancers and on the feasibility of using G4-targeting drugs in therapies. These kinds of G4 structures and their interactions with associated proteins represent a rich area of continued research, with the promise to uncover

new knowledge regarding cancer and its treatment [1].

Methodology

The several selected proteins play a critical role in the conservation of the DNA structure, especially when interacting with G4 during DNA replication and repair. For example, BRCA1 and FANCJ are engaged in the DNA repair mediated pathways which are essential for negation of malignant change [18].

Several proteins such as Nucleolin and HNRNPA1 are shown to be involved in transcription regulation by modulating the RNA polymerase activity and the formation of the transcriptional complexes at G4 sites. These interactions can influence transcriptional pausing factors, the rate of the transcription and the fidelity of the template which in turn determines the number of transcripts synthesized and the functions mapped on it signaling pathways involved in cell cycle alteration and stress responses in cells [19].

These proteins and their association with G4s are rather crucial in cancer as they imply the destabilization of the genome and changes in the expression levels of oncogenes and tumor suppressor genes. For instance, the WRN helicase is well documented to unwind G4 structures to avoid genomic instability; a characteristic of cancer cells [20].

By performing a computational analysis of these G4-binding proteins, the detailed comprehension of these proteins' ability to identify, interact, and regulate G4 structures can be achieved. This understanding is important when trying to compare and distinguish their roles in healthy human cells and information about their function in pathologies, like cancer [17].

The general goal of applying UniProt in sequence alignment is to find homologous sequence among the chosen proteins within different species. Thus, homologous sequences are the sequences that come from a common ancestor and can have structural and functional similarities if the different evolutionary branches have preserved the initial traits [21].

To conduct the sequence alignment and conservation analysis of G4-binding proteins, I utilized the UniProt database (https://www.uniprot.org/), a comprehensive and highly curated protein knowledgebase. For each protein of interest—specifically Nucleolin, HNRNPA1, BRCA1, WRN, DHX36, and FANCJ—I began by entering the protein name into the UniProt search bar. The search results often contained multiple isoforms and entries, so I selected the reviewed (Swiss-Prot) entries for Homo sapiens to ensure I was using the most validated and complete data available. From each selected human protein entry, I navigated to the "Cross-references" and "Phylogenomic databases" sections, where I located orthologous sequences for the corresponding proteins in Mus musculus (mouse), Gallus gallus (chicken), and Danio rerio (zebrafish).

Next, I downloaded the FASTA sequences for each ortholog and imported them into the UniProt alignment tool. When necessary, I used external alignment platforms such as Clustal Omega and Jalview for enhanced visualization and alignment optimization. These tools allowed for multiple sequence alignment (MSA), which revealed conserved and variable regions across species. Following alignment, I generated a percent identity matrix to quantify the similarity of each protein between species. In addition to the identity matrix, I constructed phylogenetic trees using the neighbor-joining method to examine evolutionary distances and relationships among the selected organisms. This step provided insights into how closely related these proteins are across different vertebrate classes, and where evolutionary divergence may have influenced protein function.

For a deeper understanding of mutation sensitivity and evolutionary constraint, I employed the Missense Tolerance Ratio (MTR) Viewer provided by the Regeneron Genetics Center (https://mtr-viewer.regeneron.com/). This tool aggregates genomic data to assess the tolerance of specific gene regions to missense mutations. For each protein-coding gene, I entered the gene symbol (e.g., WRN or HNRNPA1) into the search interface. The MTR Viewer then generated a plot showing missense tolerance across the entire coding sequence. Low MTR values indicated regions under strong purifying selection, suggesting these areas are less tolerant to change and likely play critical functional roles.

The Missense Tolerance Ratio was computed using a sliding window of 31 codons for the MTR analysis, as in the Regeneron MTR Viewer default parameters. The statistical stability was maintained while this window size was sufficient to identify localized regions of constraint. An MTR threshold of ¡0.3 was interpreted as regions under strong purifying selection and used to interpret conservation significance. These values were considered biologically relevant and suggestive of functional importance. Gaps and indels were removed from the MTR scoring process during alignment preprocessing to avoid the misinterpretation of mutational constraint. Thus, alignment artifacts were controlled, and only clean, high confidence coding regions contributed to the MTR based intolerance mapping.

I then correlated the low-MTR regions with known protein domains from UniProt, such as the helicase domain in WRN and the RNA-recognition motifs in HNRNPA1. By overlaying MTR data with structural domain information, I was able to identify hotspots of evolutionary conservation and infer their functional importance in genome maintenance and G4 binding. Screenshots from the MTR Viewer and relevant values were documented to support the analysis in the results section. This process allowed for a multi-dimensional evaluation of the proteins, combining evolutionary sequence conservation with empirical data on functional sensitivity to mutation.

Together, these bioinformatics tools enabled the mapping of conserved amino acid regions, identification of functionally intolerant segments, and contextualization of protein evolution across species. These analyses laid the foundation for proposing

specific domains of G4-binding proteins as critical for genomic integrity and potential therapeutic targets in oncology.

All four species were evaluated on a standardized format for each G4-binding protein to improve clarity and consistency. Specifically, for each protein, the percent identity between species was recorded, MTR profiles were interpreted within the context of functional domains and any known disease associated mutational hotspots were highlighted. To highlight the table below easily, I have broken down these findings into the table below which will make it easy for a quick glance and side by side comparison:

Protein names such as Nucleolin, HNRNPA1, BRCA1, WRN, DHX36, and FANCJ were individually searched, and their human entries were selected based on reviewed Swiss-Prot status to ensure high annotation quality. The orthologs from mouse, chicken, and zebrafish were identified through the cross-references and orthologs sections within each UniProt protein page. FASTA sequences for all species were downloaded and input into UniProt's built-in alignment tool or into external tools like Clustal Omega to perform multiple sequence alignments. A percent identity matrix was generated to assess the degree of conservation, and phylogenetic trees were constructed using neighbor-joining methods to understand evolutionary distances among species.

For analysis of missense mutation tolerance, the Regeneron Genetics Center's MTR Viewer (https://mtr-viewer.regeneron.com/) was utilized. Gene symbols for each protein of interest were entered into the search bar to retrieve MTR plots. These plots displayed regions of the gene under strong purifying selection, indicated by low MTR scores, which signify intolerance to variation. These regions were interpreted in conjunction with known functional domains, such as helicase motifs or RNA-binding regions, to determine the potential impact of mutations. Screenshots and interpretations from the MTR Viewer were incorporated into the results to highlight hotspots of evolutionary constraint, offering insights into protein regions likely essential for G4 binding and genome maintenance.

For statistical robustness of sequence conservation analyses, the multiple sequence alignment (MSA) Clustal Omega alignment quality scores were reviewed for each MSA and only alignments with quality scores greater than 85% were retained to ensure high quality comparison. When interpreting regional constraint for MTR analysis, the Regeneron MTR Viewer reported confidence intervals and regions with low MTR values and narrow confidence intervals were flagged as strongly conserved and possibly functionally indispensable. Additionally, MTR values were compared across defined domains of each protein using one way ANOVA tests and brought into visualization as bar graphs with 95% confidence intervals to show differences on missense tolerance. By supporting more rigorous evaluation the sequence constraint and evolutionary conservation, these combined statistical tools were used.

Thus, computational MTR analysis findings were cross referenced with experimental data from the literature to support and validate the results. For instance, low MTR regions in BRCA1 (600–800 AA) correspond to the BRCT domain where BRCA1 c.5266dupC (5382insC) mutations that were experimentally shown to abolish DNA repair signaling and predispose to breast and ovarian cancer occur[22]. Like this, the region of low MTR scores in the helicase domain (700–800 AA) of WRN, experimentally validated mutational hotspots that interfere with the ATPase activity, lead to chromosomal instability, and the Werner syndrome phenotypes, overlaps with WRN. Experimental studies of DHX36 have shown that mutations of the DEAH-box domain reduce G-quadruplex resolving activity, which affects immune signaling pathways[1]. Importantly, these examples of domains predicted by MTR to be the most evolutionarily conserved correlate with other domains that were previously validated in wetlab studies, which gives further support to the biological relevance of the computational results.

This integrated approach combining UniProt and Regeneron tools enabled a detailed examination of protein conservation and functional significance, helping to identify biologically important regions across species. of both sequence conservation and functional sensitivity to mutation across species for each G4-binding protein investigated.

Results

Analyzing percent identity matrix of Nucleolin proteins of different species of animals the maximum sequence similarity is evident in chicken and human Nucleolin (66.21%) and human and mouse Nucleolin (83.98%). This reveals that the functionally important domains are evolutionarily conserved across different species. These findings are supported by the phylogenetic tree, which shows that the evolutionary distances between these species are smaller, indicating that their Nucleolin proteins are more closely related.

To corroborate with these findings, the percent identity was calculated using a pairwise comparison algorithm in Clustal Omega, which aligns homologous regions across species and determines percent identity as a percentage of identical residues within the aligned region vs the alignment length. Having values of identity higher than 80% was considered strong conservation, whereas values below 60% were interpreted as potential functional divergence. Furthermore, phylogenetic trees were constructed using the neighbor joining method and bootstrap analysis was made with 1000 replicates to check the reliability of each branch. Statistically significant bootstrap values higher than 70% were considered to have strong support for the evolutionary relationships suggested among the species. To statistically validate the observed sequence conservation across different proteins and species, percent identity values were used

as the dependent variable and protein type as the independent grouping factor, and a one-way ANOVA test was performed. The p value from the ANOVA was less than 0.01, which indicates that the differences in conservation levels of the studied proteins are statistically significant and not a result of random variation. The statistical metrics of these patterns support the credibility of the observed conservation and divergence between human, mouse, chicken, and zebrafish G4 binding proteins.

Amino acid substitutions of even minor size— particularly located in functional domains such as RNA recognition motifs (RRM) can influence G4 binding affinity, affect RNA splice patterns and affect protein stability. Future studies could explore these possibilities by adding something such as structural modeling, eg AlphaFold, to predict conformational changes brought about by sequence variation. Experimental studies aimed at the use of site directed mutagenesis or binding affinity assays to define whether highly conserved proteins have full functional equivalence or have variously diverged degrees of divergence that affect G4 related activity would be also helpful.

The MTR plots show the number of genes and species that are tolerant of missense mutations within the Nucleolin protein. While comparing the susceptibility of the regions around 300 AA and 500 AA across different species, it is also observed to be higher in zebrafish rather than in other species that the amino acid at this position substitutes and have a different basic and functional requirement for species-specific Nucleolin.

In our overall comparison, which consists of proteins of humans as well as mice including HNRNPA1 as one of the proteins used, we found that there is an extraordinary level of identity or resemblance that is 99.69% which points out that the proteins are probably playing a vital role in the body and hence their sequences are conserved.

In terms of sequence conservation, the changes show a different order in the case of BRCA1 protein. It is rather surprising that the proteins encoded by human and mouse BRCA1 gene have 56.71% similarity and thus signifies that there is considerable divergence between the evolution of these species regarding suppression of cancer-related diseases.

The same observation is made by the phylogenetic analysis, in which the Nucleolin proteins of human and mouse are more closely related than the chicken and zebrafish Nucleolin. This is consistent with the typical phylogenetic tree where it is expected that the mammals are more related to each other than to birds and fish because the two classifications diverged only recently from a common group.

Merging of MTR data with the phylogenetic background makes it possible to uncover different aspects and patterns of the protein evolution process. Such areas as low MTR areas including the range of 600-800 AA in human BRCA1, demonstrate that they correspond to domains that are predominantly important for DNA repair functions. This segment is overlapping with the BRCT domain, a well characterized phosphoprotein binding motif that is involved in DNA damage response signaling. BRCA1 mutations within the BRCT domain are commonly found in hereditary breast and ovarian cancers and have been demonstrated to impair BRCA1's capacity to recruit repair factors to double stranded breaks. This reveals that alignment of these mutational hotspots with MTR scores and regions of low MTR scores are indicative that even though evolutionary constraint is related to functional necessity and disease susceptibility, there is a tighter connection. These regions are conserved differently in mice, which could mean that the mice use different ways to suppress the growth of the tumor or to repair the damage done to DNA.

The WRN proteins are highly divergent from each other across species, with a 55.11% identity between human and mouse. The most striking result was that 55.35% of zebrafish also map back to chicken, indicating potential conservation unique to-non-mammalian vertebrates.

High sequence identities between human and mouse for DHX36 (92.41%), demonstrating becoming conserved due to its functions in RNA metabolism or immune response modulation, given DHX36's involvement in recognizing and resolving G4 structures.

As HNRNPA1 is involved in splicing and transport of mRNA this high degree of conservation, particularly between human and mouse (100%), was anticipated due to its pivotal function. This trend is even more pronounced when comparing the mammals to chicken, with an average sequence identity of 81%, indicative of evolutionary adaptations in avian species. Avian species are a group of warm-blooded vertebrates constituting the class Aves, they are characterized by feather, toothless beak, laying of hard-shelled eggs, high metabolic rate, four-chambered heart and a strong but lightweight skeleton[23].

WRN had multiple regions that were highly intolerant to missense mutation throughout the entire MTR profile, but particularly in a few central domains from around 700-800 AA in both humans and mice (which correspond with known functional helicase activity motifs). The ATP binding and hydrolysis activity of the helicase domain of WRN is structurally responsible for the unwinding of DNA during replication and repair. This region has been clinically associated with mutations in Werner Syndrome and increased genomic instability in a variety of cancers, and thus is of biomedical relevance. Even single amino acid changes in the low MTR domain could disrupt WRN's ability to resolve G4 structures and maintain genome integrity under replication stress and there is overlap of this critical enzymatic region with the low MTR domain. These constrained regions are indicative of evolutionary pressure to maintain helicase functionality, crucial for DNA repair and genomic stability.

DHX36 exhibits a diverse MTR throughout species and is highly tolerant in the N-terminal regions, indicating functional versatility. The C-terminal, on the other hand, at which ATP binds and hydrolysis occurs implicates a narrower degree of

amino acid tolerance, indicating how critical an enzymatic function this is for all species. Notably, the C-terminal region is aligned with the DEAH-box helicase domain, which is required for the resolution of G4s and other structured RNAs. These variants have been linked to immune dysregulation, susceptibility to infection, because they are not able to 'sense' RNA. Low MTR values here are consistent with this region being under strong purifying selection to preserve innate immune functions and G4 resolution in RNA and DNA contexts.

We thus compared proteins in which the conservation patterns seen in G4 binding proteins were observed with homologs of homologous recombination (RAD51) and replication fork stabilization (PCNA, MRE11) proteins. Sequence alignment of human orthologs to mouse, chicken and zebrafish orthologs revealed that RAD51 and PCNA also have moderate conservation (70–85%) but proteins such as HNRNPA1 and DHX36 have obviously higher degree of sequence identity and missense intolerance. In addition, the MTR profiles of G4 binding proteins were more pronounced in regions of purifying selection than the replication associated controls. Thus, G4-binding proteins would be subject to stronger evolutionary constraints (probably due to playing a central role in handling of secondary DNA structures, and replication fidelity under stress). These results are consistent with the conjectured that the observed conservation patterns of G4 binding proteins are not universal properties of genome maintenance proteins, but rather reflect specialized selective pressures associated to a particular interaction with G4 motifs and replication stress linked to cancer.

HNRNPA1 exhibits high missense mutations tolerance; however, the RNA recognition motifs (RRMs) are a key discrepancy in this pattern, particularly RRM1 and RRM2, which show lower tolerance in humans and mice. This pattern underscores the importance of these domains in RNA-binding affinity and specificity, critical for post-transcriptional regulation.

Conclusion

G4s are structures within nucleic acids constituted by regions of high guanine content stacking in four-stranded arrangements. In the last twenty years G4s have been investigated as molecular controllers that can be involved in different important processes of the cell life such as DNA replication, gene transcription and cell aging. Most importantly, G4s have attracted considerable interest due to their possible involvement in the tumor advancement and development, and thus, can be a focus of more novel anti-cancer treatments. However, as contradicted to much established facts that relate G4s to disease progression, nonetheless, very definitive proof that connects these structures with cancer initiation is still missing.

The current study examined the involvement of G4s in DNA replication as well as cancer cell proliferation. It was established that when G4s form in DNA, then they naturally arrest the DNA replication forks, and thus results in replication stress, which prevails in cancer cells. In the process of transcription G4s regulate cell growth and differentiation through effecting the levels of oncogenes and tumor repressor genes in the cancer cells. Also, the fact that G4s are comprised of four coordinated guanines indicates their participation in cellular aging by interacting with telomeres and their functionalities. This leads to oncogenesis, and cellular senescence, a state where cells stop dividing but do not die it.

The identified DNA structure called G4 has been shown to be critical for DNA replication as well as proliferation of cancer cells due to functions as molecular barrier control in critical cellular activities. Due to their amenities to generate replication stress and manipulate gene manifestation, they remain prominent in cancer biology. In addition, G4s connect with the process of cellular aging and telomere stability with carcinogenesis and cellular senescence. Thus, G4s have emerged as promising G4s as targets capable of inhibiting the proliferation of cancer cells and stabilizing genomes. Subsequent research should try to generate more compelling data about their relationships to the actual development of cancer and about the possibility of targeting G4 formation and integrity for those purposes.

Through the control of G4 in oncogene transcriptional activity, G4s can accelerate oncogenesis. G4's motifs are often found in the promoter regions of proto-oncogenes, such as MYC and KRAS. Proto-oncogenes are normal genes that play a crucial role in cell growth, division, and other processes. They contain the necessary information for the synthesis of proteins responsible for stimulating cell division, inhibiting cell differentiation, and preventing apoptosis (cell death). However, if a mutation occurs in a proto-oncogene, it can become permanently activated, turning into a malfunctioning gene called an oncogene. This can lead to uncontrolled cell growth, which is a defining feature of cancerous tumors. Specifically, MYC is a gene that codes for a transcription factor, regulating numerous genes involved in cell growth and division. Overexpression of MYC can lead to uncontrolled cell proliferation, contributing to oncogenesis. On the other hand, KRAS provides instructions for making a protein called K-Ras, which is part of the RAS/MAPK pathway. This protein relays signals from outside the cell to the cell's nucleus, instructing the cell to grow and divide or to mature and take on specialized functions. Stabilization of G4s in these regions can boost transcriptional activity of these genes. High transcription levels of oncogenes such as MYC and KRAS due to stabilized G4s interfere with the normal functioning of cancerous cells through the activation of critical cellular pathways, such as cell cycle, cell growth and cell survival. Upregulation of these oncogenes is one of the major driving forces for tumor initiation and progression, including abnormal and excessive cellular proliferation and cancerous growth[2]. This is linked to oncogene expression, because the G4 structures recruit and bind transcription factors and other regulatory proteins that are then

stabilized, or activated, in the presence of the G4 structure. This binding recruits the other proteins and can change the local chromatin structure so that it becomes more open to the transcription of that gene[2]. In contrast, the modulation of G4s enhancing tumor suppressor genes often reduces their expression. Formation of G4 structures within or near the promoter regions of tumor suppressor genes is the favorable pathway through which the G4s can inhibit their expression. This is because the binding of transcription machinery, and hence initiation of transcription, can be impeded if G4 structures form in these regions. Moreover, if tumor suppressor genes (genes that directs the production of a protein that is part of the system that regulates cell division) such as TP53 or RB1 are downregulated, cell cycle checkpoints as well as apoptotic machinery are removed, facilitating the unregulated growth of cells akin to that found in cancerous cells[12].

G4's can be harmful because they physically exist during DNA replication in S phase which is critical for tumor cell growth. The replication machinery can meet G4 structures during DNA replication, and encounter these structures as a barrier, halting the replication forks in its tracks. These stalls are not merely delays they are sites of genomic instability, where these stalls can lead to mutational accumulation, chromosomal breaks and chromosomal fusion or incomplete DNA replication, all of which are damaging to the integrity of the genome and can further suppress expression and or function of tumor suppressor genes[4]

Because of their central roles in promoting oncogene activity and silencing tumor suppressor genes, these G4s serve as a molecular double-edged sword in cancer biology and offers new hope regarding therapeutic interventions. For example, we may be able to modulate their formation or stability by targeting G4s with small molecules or ligands to regulate the expression of these genes, which is now at the forefront of active research. We can block the stabilization of G4 structures in oncogene promoters while enhancing their stability in the promoters of silences tumor suppressor genes to promote their protective effects[24].

Moreover, pinpointing the specificity with which G4s alter gene expression and replication provides the opportunity to develop cancer therapies that act selectively on these mechanisms. By modulating the stability of G4s or resolving their structures, it might be possible to intervene in the progression of the disease more effectively. In sum, the promising therapeutic targets offered by G4s have made them a prominent focus of future oncological strategies[25].

G4's are unusual structures formed by specific sequences of nucleic acids found frequently in gene promoter regions (it's the section of DNA that is associated with the activation or deactivation of a gene). They are abundant within the promoter regions of genes like MYC and KRAS that play a critical role in cancer development and progression. Their stabilization can lead to overexpression of these important genes – an activity often associated with oncogenic function[26]. On the other hand, G4's can also shut off important tumor suppressor genes, which can cause crucial silencing of critical tumor-protective genes that might lead to loss of cellular regulatory factors that contribute to carcinogenesis[3].

Through ensuring structural stability at both telomeric and gene promoter regions, daily G4 formation helps maintain the integrity of the genome. By preventing improper rearrangements of the genome and DNA damage, G4's protect the cellular genetic information, and reduce the chance of mutation and cancer[17].

Small-molecule compounds have been designed to target G4's to modulate them towards increasing or decreasing their stability. This strategy can use specific sequences in the cancer genome to alter how cells express their genes, opening new avenues in the search for treatments. Small molecules have the potential to stabilize G4's at oncogene promoters and inhibit their overactive expression, slowing tumor growth and spread[13].

Because G4 targeting compounds are selective, they can be used along with existing cancer drugs to potentiate their effects. By modulating the expression of genes that are crucial for cancer cell survival and proliferation, these chemicals make cancer cells more susceptible to standard anti-cancer drugs, including chemotherapy and radiation therapy[1].

This research continues to reveal new pathways and molecular mechanisms involved in G4 functions and is essential for targets and for understanding how modulating G4 stability might influence disease progression and response to anticancer treatments[2].

The study of G4's in cancer, though a young field of study, promises to transform the way that cancer is treated. Future studies aimed at understanding the organization of chromatin and how G4's are involved in gene expression regulation will likely lead to the development of highly specific cancer therapies with high efficacy and reduced toxicity[27].

The pivotal roles of G4s in the replication of DNA and the cell proliferation of cancer cells have been thoroughly documented, highlighting numerous features of different aspects of G4 interactions, cellular mechanisms and therapeutic avenues[9].

Sequence identities of 99.69% HNRNPA1 (human) versus HNRNPA1 (mouse) and 92.41% DHX36 (human) versus HNRNPA1 (mouse) indicate the importance of these two proteins in G4 recognition and RNA processing. However, the MTR profiles, and specifically the number of MTR states in as much as 100 folded degrees of freedom in the bacterial WRN and mammalian BRCA1 proteins, as well as the fact that they have lower identity across the species, indicate functional divergence in DNA repair mechanisms. The conserved domains, especially those with low missense tolerance, are likely to be essential functional regions, and mutations in these regions may cause genomic instability and thus contribute to cancer development.

The discussion section summarizes results; it also discusses their context by relating them to the previous research and discusses its implications. For example, the extreme high sequence identity (e.g. 99.69% HNRNPA1, 92.41% DHX36 for human and mouse) for HNRNPA1 and DHX36 both conserved and essential roles in RNA processing and G4 resolution. The high degree of conservation is important because it implies that these proteins have important biological roles that have been conserved through evolution because of strong functional constraints. Such conservation in evolutionary biology usually indicates that changes in these regions would be deleterious, thus these proteins are essential for maintaining core cellular processes between species. It is likely that these proteins have central roles in keeping transcriptome stable and resolving replication stress, especially under conditions that promote cancer. On the other hand, compared to proteins such as BRCA1 and WRN, whose identity was low across species and MTR profiles variable, these proteins are likely to play more specific or divergent roles in DNA repair across organisms. Notably, the WRN helicase showed extreme intolerance to missense mutations both in the helicase domain after an N97Q truncation within the 300–500 amino acid range and in the 700–800 amino acid range, suggesting the functional importance of retaining that helicase domain to maintain genome integrity. This implies that structure and evolutionarily selected stability and function of G4-protein interactions. Future research could be done to see how disrupting one of the low-MTR regions impacts DNA repair fidelity and cancer progression through functional assays. In addition, mapping protein–G4 interaction sites on a higher resolution across species may reveal species specific targets as opposed to doing so for future therapeutic development. In turn, these findings shed light on the conservation and variability of the molecular factors that bind G4 and create the possibility to use precision medicine focused on seeking out proteins to which one can target G4 based on their conserved domains.

References

1 D. Rhodes and H. Lipps, *G-quadruplexes and their regulatory roles in biology*.

2 N. Kosiol, S. Juranek, P. Brossart, A. Heine and K. Paeschke, *G-quadruplexes: a promising target for cancer therapy*.

3 J. Figueiredo, J. Mergny and C. Cruz, *G-quadruplex ligands in cancer therapy: Progress, challenges, and clinical perspectives*.

4 M. Eppard, J. Passos and S. Victorelli, *Telomeres, cellular senescence, and aging: past and future*.

5 H. Shu, R. Zhang, K. Xiao, J. Yang and X. Sun, *G-Quadruplex-Binding Proteins: Promising Targets for Drug Design*, https://doi.org/10.3390/biom12050648.

6 J. Shay and W. Wright, *Senescence and immortalization: role of telomeres and telomerase*.

7 K. Wang, Y. Wang, J. Dickerhoff and D. Yang, *DNA G-quadruplexes as targets for natural product drug discovery*, https://doi.org/10.1016/j.eng.2024.03.015.

8 M. Finch-Edmondson and M. Sudol, *Framework to function: mechanosensitive regulators of gene transcription*.

9 A. Magis, S. Manzo, M. Russo, J. Marinello, R. Morigi, O. Sordet and G. Capranico, *DNA damage and genome instability by G-quadruplex ligands are mediated by R loops in human cancer cells*.

10 J. Robinson, G. Flint, I. Garner, S. Galli, T. Maher, M. Kuimova, R. Vilar, I. McNeish, R. Brown, H. Keun and M. Antonio, *G-quadruplex structures regulate long-range transcriptional reprogramming to promote drug resistance in ovarian cancer*, https://doi.org/10.1101/2024.06.24.600010.

11 H. Masai and T. Tanaka, *G-quadruplex DNA and RNA: Their roles in regulation of DNA replication and other biological functions*.

12 J. Figueiredo, M. Djavaheri-Mergny, L. Ferret, J. Mergny and C. Cruz, *Harnessing G-quadruplex ligands for lung cancer treatment: A comprehensive overview*.

13 A. Criscuolo, E. Napolitano, C. Riccardi, D. Musumeci, C. Platella and D. Montesarchio, *Insights into the Small Molecule Targeting of Biologically Relevant G-Quadruplexes: An Overview of NMR and Crystal Structures*, https://doi.org/10.3390/pharmaceutics14112361.

14 B. Bahls, I. Aljnadi, R. Emídio, E. Mendes and A. Paulo, *G-Quadruplexes in C-MYC promoter as targets for cancer therapy*, https://doi.org/10.3390/biomedicines11030969.

15 A. Henderson, Y. Wu, Y. Huang, E. Chavez, J. Platt, F. Johnson, R. Brosh, D. Sen and P. Lansdorp, *Detection of G-quadruplex DNA in mammalian cells*.

16 Z. Zhang, S. Qian, D. Wei and Z. Chen, *In vivo dynamics and regulation of DNA G-quadruplex structures in mammals*.

17 V. Sanchez-Martin, *DNA G-Quadruplex-Binding Proteins: An updated overview*, https://doi.org/10.3390/dna3010001.

18 C. Bradley, *A statistical framework for rare disease diagnosis*.

19 D. Day, B. Zhang, S. Stevens, F. Ferrari, E. Larschan, P. Park and W. Pu, *Comprehensive analysis of promoter-proximal RNA polymerase II pausing across mammalian cell types*.

20 W. Tang, A. Robles, R. Beyer, L. Gray, G. Nguyen, J. Oshima, N. Maizels, C. Harris and R. Monnat, *The Werner syndrome RECQ helicase targets G4 DNA in human cells to modulate transcription*.

21 G. Mayr, F. Domingues and P. Lackner, *Comparative analysis of protein structure alignments*.

22 M. Silk, S. Petrovski and D. Ascher, *MTR-Viewer: identifying regions within genes under purifying selection*.

23 K. Idahor, *Avian reproduction*, https://doi.org/10.5772/intechopen.101185.

24 C. Nakanishi and H. Seimiya, *G-quadruplex in cancer biology and drug discovery*.

25 Y. Wang, J. Yang, A. Wild, W. Wu, R. Shah, C. Danussi, G. Riggins, K. Kannan, E. Sulman, T. Chan and J. Huse, *G-quadruplex DNA drives genomic instability and represents a targetable molecular abnormality in ATRX-deficient malignant glioma*.

26 S. Balasubramanian, L. Hurley and S. Neidle, *Targeting G-quadruplexes in gene promoters: a novel anticancer strategy?*

27 T. Richl, J. Kuper and C. Kisker, *G-quadruplex-mediated genomic instability drives SNVs in cancer*, https://doi.org/10.1093/nar/gkae098.