G4-associated human diseases

Nancy Maizels*

Abstract

Recent research has established clear connections between G-quadruplexes and human disease. Features of quadruplex structures that promote genomic instability have been determined. Quadruplexes have been identified as transcriptional, translational and epigenetic regulatory targets of factors associated with human genetic disease. An expandable GGGGCC motif that can adopt a G4 structure, located in the previously obscure C9ORF72 locus, has been shown to contribute to two well-recognized neurodegenerative diseases, amyotrophic lateral sclerosis (ALS) and frontotemporal dementia (FTD). This review focuses on these advances, which further dispel the view that genomic biology is limited to the confines of the canonical B-form DNA duplex, and show how quadruplexes contribute spatial and temporal dimensionalities to linear sequence information. This recent progress also has clear practical ramifications, as prevention, diagnosis, and treatment of disease depend on understanding the underlying mechanisms.

Keywords ALS; FTD; G4; helicase; quadruplex

Introduction

G-quadruplex or “G4” structures form in G-rich regions of single-stranded DNA or RNA [1,2]. Quadruplexes are distinct in structure from the canonical B-form duplex, and they are recognized by a distinct subset of nuclear factors. Formation of quadruplexes can be regulated, since it occurs only in single-stranded regions, which are generated during replication and transcription as a result of high density supercoiling. This review discusses some recent advances in our understanding of the many ways in which genomic biology exploits the special properties of quadruplexes, and how quadruplexes are involved in human health and disease.

G4 motifs and quadruplex structures

A G4 motif is a sequence that can form a quadruplex structure (Sidebar A). The G4 signature motif is comprised of four adjacent runs of guanines separated by other nucleotides: G$_{3}$N$_{x}$G$_{3}$N$_{x}$G$_{3}$N$_{x}$G$_{3}$N$_{x}$. The recurrent unit in a quadruplex is a G-quartet [3], a planar array of four hydrogen-bonded guanines (Fig 1A). The guanines in the G-runs in the G4 sequence motif typically participate in G-quartet formation. Strands in parallel or antiparallel orientation may contribute guanines to a G-quartet (Fig 1B). The planar quartets are connected by “loops” formed by the nucleotides between the G-runs, denoted as “N” in the G4 sequence motif.

The G$_{3}$N$_{x}$G$_{3}$N$_{x}$G$_{3}$N$_{x}$G$_{3}$ sequence motif constitutes a loose rather than strict consensus. Some sequences that do not strictly conform to the consensus do form quadruplexes—CGG repeats are one example. Conversely, about 10% of oligonucleotides bearing sequences that conform to that motif do not readily form quadruplex structures in vitro, although it remains unclear whether the failure of some G4 motifs to form quadruplexes reflects flaws in the algorithm or in the conditions used to test structure formation.

Quadraplex structures are diverse, and they may be quite intricate [4,5]. Quadruplex conformation is determined by the orientation of the glycosidic bonds of the guanines in the quartets, the parallel, anti-parallel, or mixed orientation of the strands contributing guanine bases to the quartets, and the lengths and sequences of the loops connecting the guanine runs (Fig 1B). Quadruplexes exhibit considerable thermodynamic stability, largely derived from stacking of the hydrophobic G-quartets, and modulated by the length and sequence of the loops. It is not yet possible to predict

(continued)
transcribed from telomeric repeats, quadruplexes composed of parallel structures as DNA or RNA [10,11]. In TERRA RNAs the human TTAGGG telomeric repeat, which forms characteristic can fold to produce regular structures. The best-studied example is where abundant G4 motifs form well-defined structures that telomere, with its clear connection to cell proliferation and aging, remains high. One target of continued therapeutic interest is the G4 motifs in human cells [8]. The interest in developing G4 ligands should in principle be good targets for G4 ligands [9].

Evidence that distinct G4 motifs adopt distinctive quadruplex structures fueled efforts to develop small molecule ligands that would bind to a postulated quadruplex and affect expression of a nearby gene [6]. This strategy has not yet generated any drugs that can be used in the clinic, but G4 ligands have proved useful for other purposes. Notably, the G4 ligand PhenDC3 has been used to exacerbate instability at G4 motifs in S. cerevisiae [7], and a different compound, pyridostatin, has been shown to cause instability at G4 motifs in human cells [6]. This strategy has not yet generated any drugs that can be used in the clinic, but G4 ligands have proved useful for other purposes. Notably, the G4 ligand PhenDC3 has been used to exacerbate instability at G4 motifs in S. cerevisiae [7], and a different compound, pyridostatin, has been shown to cause instability at G4 motifs in human cells [6].

G4 motifs exhibit a distinct distribution within the human genome and among human genes [17]. As G4 motifs can confer or correlate with specific properties and functions, it is useful to distinguish regions that are significantly enriched for G4 motifs against the genomic average as “G4hi” and “G4lo”, respectively. G4 motifs are enriched in three G4hi chromosomal domains: the telomeres, composed of G4 motifs in a genome can be tallied and mapped by algorithms that search for matches to the G4 sequence motif, $3\text{NxG}$, which forms a parallel quadruplex in which r(UUAGGG)$_4$ repeats readily stack on each other at their 5' ends (Fig 1C; [12]). The 52 nt human CEB25 VNTR, located on chromosome 10, contains a 26 nt core sequence, d(AAGGGTGGTTGTAAGTGGGGTGGGTG), which forms a parallel quadruplex in which a 9 nt loop connects the top of one G-quartet to the bottom of another; and in long repeats, a 26 nt chain of relatively unstructured DNA connects these quadruplexes to create a “pearl necklace” conformation ([13]; Fig 1D). The quartets, the loops within them, and the chains that connect them are all potential targets for protein binding.

A number of factors have been shown to bind quadruplexes in human cells, and some recurrent polypeptide domains have been shown to promote binding to quadruplexes. These include the RRM/RBD domains and RGG motifs found in many nuclear RNA binding proteins [14], the RSM domain found at the N-terminus of RHAU RNA helicase [15], and the RQC domain found in most RECQ family helicases [16]. Loops within quadruplexes may confer specificity to protein binding. Some of the factors that interact with quadruplexes will be discussed later in this review.

G4 motifs in the genome

We have proposed that G4 motifs and quadruplex structures that participate in key processes comprise the G4 genome, analogous to the transcriptome, proteome, or metabolome [17]. This poses the challenge of identifying those G4 motifs that do form quadruplexes in living cells, and determining their specific functions.

G4 motifs in a genome can be tallied and mapped by algorithms that search for matches to the G4 sequence motif, $3\text{NxG}$, which forms a parallel quadruplex in which r(UUAGGG)$_4$ repeats readily stack on each other at their 5' ends (Fig 1C; [12]). The 52 nt human CEB25 VNTR, located on chromosome 10, contains a 26 nt core sequence, d(AAGGGTGGTTGTAAGTGGGGTGGGTG), which forms a parallel quadruplex in which a 9 nt loop connects the top of one G-quartet to the bottom of another; and in long repeats, a 26 nt chain of relatively unstructured DNA connects these quadruplexes to create a “pearl necklace” conformation ([13]; Fig 1D). The quartets, the loops within them, and the chains that connect them are all potential targets for protein binding.

A number of factors have been shown to bind quadruplexes in human cells, and some recurrent polypeptide domains have been shown to promote binding to quadruplexes. These include the RRM/RBD domains and RGG motifs found in many nuclear RNA binding proteins [14], the RSM domain found at the N-terminus of RHAU RNA helicase [15], and the RQC domain found in most RECQ family helicases [16]. Loops within quadruplexes may confer specificity to protein binding. Some of the factors that interact with quadruplexes will be discussed later in this review.

G4 motifs in the genome

We have proposed that G4 motifs and quadruplex structures that participate in key processes comprise the G4 genome, analogous to the transcriptome, proteome, or metabolome [17]. This poses the challenge of identifying those G4 motifs that do form quadruplexes in living cells, and determining their specific functions.

G4 motifs in a genome can be tallied and mapped by algorithms that search for matches to the G4 sequence motif, $3\text{NxG}$, which forms a parallel quadruplex in which r(UUAGGG)$_4$ repeats readily stack on each other at their 5' ends (Fig 1C; [12]). The 52 nt human CEB25 VNTR, located on chromosome 10, contains a 26 nt core sequence, d(AAGGGTGGTTGTAAGTGGGGTGGGTG), which forms a parallel quadruplex in which a 9 nt loop connects the top of one G-quartet to the bottom of another; and in long repeats, a 26 nt chain of relatively unstructured DNA connects these quadruplexes to create a “pearl necklace” conformation ([13]; Fig 1D). The quartets, the loops within them, and the chains that connect them are all potential targets for protein binding.

A number of factors have been shown to bind quadruplexes in human cells, and some recurrent polypeptide domains have been shown to promote binding to quadruplexes. These include the RRM/RBD domains and RGG motifs found in many nuclear RNA binding proteins [14], the RSM domain found at the N-terminus of RHAU RNA helicase [15], and the RQC domain found in most RECQ family helicases [16]. Loops within quadruplexes may confer specificity to protein binding. Some of the factors that interact with quadruplexes will be discussed later in this review.

G4 motifs exhibit a distinct distribution within the human genome and among human genes [17]. As G4 motifs can confer or correlate with specific properties and functions, it is useful to distinguish regions that are significantly enriched for G4 motifs against the genomic average as “G4hi” and “G4lo”, respectively. G4 motifs are enriched in three G4hi chromosomal domains: the telomeres, composed of 3–20 kb TTAGGGG repeats; the rDNA, where G4 motifs are abundant in sequences transcribed into rRNA as well as spacer DNA; and the immunoglobulin heavy chain switch regions, 1–10 kb degenerate repetitive regions that are sites of class switch recombination junctions. Like CEB25 (see above), many minisatellites or VNTRs contain G4 motifs.

Among the approximately 30,000 human RefSeq genes, promoters are typically G4hi, especially in the 1 kb region flanking the TSS; coding regions are G4lo; and G4 motifs are greatly enriched at the coding regions, 107 vs. 3.2 × 10⁶ bp), has only 0.01% as many G4 motifs. This suggests that the functions of G4 motifs have changed over the course of evolution.

G4 motifs exhibit a distinct distribution within the human genome and among human genes [17]. As G4 motifs can confer or correlate with specific properties and functions, it is useful to distinguish regions that are significantly enriched for G4 motifs against the genomic average as “G4hi” and “G4lo”, respectively. G4 motifs are enriched in three G4hi chromosomal domains: the telomeres, composed of 3–20 kb TTAGGGG repeats; the rDNA, where G4 motifs are abundant in sequences transcribed into rRNA as well as spacer DNA; and the immunoglobulin heavy chain switch regions, 1–10 kb degenerate repetitive regions that are sites of class switch recombination junctions. Like CEB25 (see above), many minisatellites or VNTRs contain G4 motifs.

Among the approximately 30,000 human RefSeq genes, promoters are typically G4hi, especially in the 1 kb region flanking the TSS; coding regions are G4lo; and G4 motifs are greatly enriched at the coding regions, 107 vs. 3.2 × 10⁶ bp), has only 0.01% as many G4 motifs. This suggests that the functions of G4 motifs have changed over the course of evolution.
development, and they are enriched for G4 motifs in both exons and introns. Conversely, tumor suppressor genes are typically G4lo.

As discussed below, G4 motifs may influence genomic stability, transcriptional and translational regulation, and chromatin structure, so enrichment or depletion of G4 motifs may confer special properties or regulation.

Quadruplexes add a temporal dimension to genomic structure and function

The human genome contains many G4 motifs: over 360,000 when searched using an algorithm that allows up to 7 nt in the loops between G-runs, and approximately twice that number allowing...
loops up to 12 nt in length. The fraction of genomic G4 motifs that actually form quadruplex structures in a living human cell is not known.

Quadruplexes can form in regions of B-form duplex DNA that are denatured by replication or transcription, or in which local superhelicity has been altered (Fig 2). Quadruplex density within DNA molecules displays striking dependence on negative superhelicity [21], which increases in the wake of the advancing transcription apparatus, and can also increase in response to topoisomerase activity or protein binding. Conversely, a quadruplex can stall replication or transcription, or alter DNA structure to inhibit recognition of the duplex by a sequence-specific DNA binding protein (Fig 2). Thus, quadruplex formation is regulated and can respond to temporal signals associated with the cell cycle and development.

Tools that detect, enumerate and map quadruplexes would be very useful for identifying those G4 motifs that do indeed form quadruplex structures, and establishing the conditions that promote or prevent quadruplex formation. Antibodies have been reported that bind quadruplexes specifically in vitro, and that stain living cells (reviewed by [22]). If those antibodies were able to recognize all quadruplex structures, then it would in principle be possible to generate a catalog of quadruplexes by ChIP-Seq. However, no anti-G4 antibody has yet been shown to recover G4 motifs as a specific or even predominant target by genome-wide approaches. Some efforts are directed toward the development of ligands that light up upon quadruplex binding [23], but these have not reached the stage where they could be used to enumerate quadruplex structures.

A more fruitful approach to identification of quadruplexes in a living cell is to map sites of association of endogenous quadruplex binding proteins. ChIP-Seq analysis has shown that G4 motifs are greatly overrepresented (> 40%; P = 0) among sites of association of two endogenous G4 helicases, XPB and XPD, and that 2% of G4 motifs associated with these factors (P = 0; [24]). Those experiments depended on the availability of highly specific antibodies to XPB and XPD, which thus far are the only quadruplex binding proteins for which genome-wide ChIP-Seq has been carried out. It has become possible to use CRISPR/Cas9 to add an epitope tag to essentially any gene, circumventing the need for specific antibodies for ChIP-Seq. This means that it is now possible in principle to catalog the nuclear DNA targets of all endogenous quadruplex binding proteins and provide a more complete picture of the “G4 Genome”.

Figure 2. Induction of quadruplex structures and potential impact on essential processes. (A) Replication, transcription, or local increases in negative superhelicity may expose a single-stranded region bearing a G4 motif, and lead to quadruplex formation. [G4] identifies one or more G4 motifs. G4 RNA may form in the transcript of a gene carrying a G4 motif on the non-transcribed strand (center). (B) A quadruplex may stall replication or transcription, or prevent a sequence-specific transcription factor from recognizing or binding its site in duplex DNA.
Quadruplexes pose challenges for maintaining genomic stability

A novel role for Pol θ in promoting G4 motif stability in C. elegans

A G-quadruplex structure on the template strand for replication may cause the replication apparatus to stall (Fig 2A). This may in turn promote formation of one-ended breaks in the chromosome which can initiate genomic instability if not repaired faithfully [25]. Our understanding of replication-associated instability at G4 motifs has benefited from experiments carried out in model organisms, especially the nematode C. elegans and the yeast S. cerevisiae.

Replicative instability at G4 motifs was first demonstrated in experiments by Tijsterman and collaborators that analyzed DNA deletions in nematodes deficient in dog-1, an ortholog of human FANCJ [26] (reviewed by [27]). Deletion endpoints mapped to G4 motifs, and only to G4 motifs. Unexpectedly, repair at G4-induced breaks in the nematode depends on a novel repair mechanism, in which the A-family polymerase, Pol θ, tethers the end of the break and carries out error-prone repair synthesis to generate occasional insertions at the break sites [28]. The importance of Pol θ is especially evident in backgrounds lacking low fidelity polymerases Pol η and Pol κ [29] (reviewed in [30]).

Replication-associated genomic instability at the G4L1 motif in S. cerevisiae

Many DNA repair pathways are highly conserved, and S. cerevisiae has proved to be an excellent model for eukaryotic DNA repair in many contexts. Nicolas and collaborators used a human G4hi VNTR, as a reporter for replicative instability at G4 motifs, and identified Pif1 helicase as necessary for G4 motif stability [7]. Instability exhibited a clear strand bias and occurred only if the G-rich strand was the template for leading strand replication [31] (reviewed by [32]). Analysis of another G4th human VNTR, CEB25 [33], suggested that G4 motifs differ in their ability to cause replicative instability. This has now been validated by a systematic analysis from the Nicolas laboratory, showing that G4 motifs that contain single nucleotide loops (“G4L1 motifs”) are most detrimental to genomic stability [19].

G4L1 motifs are depleted in most genomes, including those of S. cerevisiae, S. pombe, nematode, and human. Depletion particularly affects G4 motifs with loops containing pyrimidines rather than purines, which exhibit the greatest genetic instability and the greatest thermodynamic stability. The entire yeast genome contains only four G4 motifs with short loops (≤ 3 nt), and even these are not likely sites of replication-associated genomic instability, as they are purine rather than pyrimidine loops. If the same rules of genomic maintenance that have been documented in S. cerevisiae extend to human cells, the encouraging take home message is that only 5% of the G4 motifs in the human genome (18,000 of 360,000) are G4L1 motifs, and thus likely to be at risk of maximal genomic instability.

The recent evidence that potentially unstable G4 motifs are dramatically depleted from the S. cerevisiae genome [19] has raised questions about a report claiming that the S. cerevisiae Pif1 helicase is enriched at G4 motifs and suppresses G4-induced genomic instability at those sites [34,35]. There were initially questions about how G4 motifs were identified in that report, as it employed a search algorithm that allowed a maximal loop length of 25 nt [36], which will considerably inflate the numbers of G4 motifs scored. A search algorithm that allows loops up to 25 nt in length tallies 552 G4 motifs in the S. cerevisiae genome, while the more standard algorithm that allows loops up to 7 nt in length tallies only 27. If the endogenous G4 motifs in S. cerevisiae are not sites of replicative instability, as suggested [19], then enrichment of Pif1 at or near a G4 motif is unlikely to reflect a role for this factor in suppressing G4-induced genomic instability.

Analysis of sites of genomic instability in nematodes and human cells provides further support for the notion that G4L1 motifs—and not G4 motifs in general—are sequences at risk [19]. The genome of the nematode C. elegans contains unusual mono-G runs, which account for essentially all of the G4L1 motifs in this organism. DNA deletions in nematodes deficient in dog-1, an ortholog of human FANCJ, map almost exclusively to mono-G runs [26]. A corresponding effect is evident in human cells treated with the G4-ligand pyridostatin, which induces genomic instability accompanied by induction of the alternate histone, γ-H2AX [8]. Sites of γ-H2AX accumulation proved to be enriched for pyrimidine-containing but not purine-containing G4L1 motifs [19].

Further experiments are necessary to better define the G4 motifs that do in fact contribute to genomic instability in human cells, and to understand which factors normally enable smooth replication of quadruplexes. A significant enrichment of G4 motifs has been documented at chromosomal breakpoints in human tumor genomes [37], and more detailed examination of loop length and composition of those motifs may provide useful rules for identifying sites of instability. In addition, as increasing numbers of normal and cancer genomes are systematically sequenced, it should be possible to make statistically compelling correlations between mutations and polymorphisms in repair and replication factors and instability at G4 motifs genome-wide. These same factors may contribute to preventing expansion of G4 motifs associated with neurodegenerative diseases, as discussed below.

G4 helicases promote stability at G4 motifs in human cells

Connections between quadruplexes and human disease (summarized in Table 1) first emerged from studies of the roles of G4 helicases at regions of the genome significantly enriched for G4 motifs. Genome-wide analyses have recently established the importance of smooth replication for the maintenance of genomic stability at G4 motifs.

BLM helicase at the immunoglobulin switch regions

BLM helicase is one of five human RecQ family helicases [38]. BLM is deficient in the genetic disease Bloom syndrome, a rare autosomal recessive disease characterized by primordial dwarfism and predisposition to cancer. Affected individuals frequently present with immunodeficiency due to impaired immunoglobulin heavy chain class switch recombination. In class switch recombination, regulated DNA deletion joins the expressed variable region to a new constant region, conferring new effector functions on the encoded antibody without changing its specificity for antigen. Recombination junctions map to intronic “switch” regions, 1–10 kb degenerative repeats that are enriched for G4 motifs. Recombination to a particular switch
region is preceded by its active transcription, which promotes quadruplex formation. BLM is a G4 helicase, which preferentially unwinds quadruplex relative to duplex DNA [39], and can thereby remove quadruplexes from the transcribed switch regions to enable replication of switched B cells and a robust immune response.

**Table 1. Human disease-associated factors that bind quadruplexes.**

<table>
<thead>
<tr>
<th>Helicase</th>
<th>Disease</th>
<th>Process</th>
</tr>
</thead>
<tbody>
<tr>
<td>XPB/XPD</td>
<td>Xeroderma pigmentosum (UV sensitivity, cancer); trichothiodystrophy and Cockayne syndrome (development)</td>
<td>TFIH components, with functions in nucleotide excision repair</td>
</tr>
<tr>
<td>BLM</td>
<td>Primordial dwarfism, genomic instability, cancer predisposition</td>
<td>DNA replication; immunoglobulin class switch recombination</td>
</tr>
<tr>
<td>WRN</td>
<td>Premature aging, cancer predisposition, cell senescence</td>
<td>Telomere maintenance</td>
</tr>
<tr>
<td>RTEL1</td>
<td>Hoyeraal-Hreidarsson syndrome (cf. dyskeratosis congenita), cell senescence</td>
<td>Telomere maintenance</td>
</tr>
<tr>
<td>DNA2</td>
<td>Cell senescence</td>
<td>Telomere maintenance; helicase and G4 nuclease</td>
</tr>
<tr>
<td>FANCJ</td>
<td>Fanconi anemia (bone marrow failure)</td>
<td>Replication with REV1 polymerase (incorporates only C)</td>
</tr>
<tr>
<td>CHL1</td>
<td>Warsaw break syndrome (cohesinopathy)</td>
<td>Chromosome segregation</td>
</tr>
<tr>
<td>ATRX</td>
<td>Alpha thalassemia with mental retardation</td>
<td>Remodels chromatin at G4 motifs genome-wide (telomeres, VNTRs, etc.)</td>
</tr>
<tr>
<td>eIF4A</td>
<td>T cell leukemia (overexpression)</td>
<td>Translation: unwinds mRNA to scan for 5’-AUG initiator codon</td>
</tr>
</tbody>
</table>

**Expansible G4 motifs in neurological diseases**

At least thirty neurological diseases are associated with expansions at DNA repeats, including G4 motifs and trinucleotide repeats (TNRs) [57].

**Fragile X syndrome**

The most recognized of the human G4 expansion diseases is Fragile X syndrome, the most common cause of inherited mental retardation, with symptoms that frequently include behavioral problems. Fragile X syndrome is a neurodevelopmental disease caused by expansions of a CGG repeat in the FMR1 gene, with disease-associated repeat numbers of more than 200 [58]. This repeat does not strictly conform to the G4 motif, as the G-runs are only two nucleotides in length, but it does form quadruplex structures. Fragile X syndrome is characterized by anticipation: disease severity increases and age of onset decreases from one generation to the next, reflecting increased repeat lengths between generations. However, CGG repeat lengths are identical among tissues of an individual affected with Fragile X syndrome, as the CGG repeat expands during cell divisions that occur in germ cells and early embryonic development, but not in fully differentiated somatic tissues. The developmental regulation of CGG repeat expansion reflects a programmed switch in replication origin use. Expansion occurs when the expandable sequence is the template for leading strand replication, reflecting structure formation within Okazaki fragments replicating the lagging strand [59,60].
Shorter CGG repeat expansions, on the order of 55–200, can cause a distinct, late onset neurodegenerative disease, Fragile X-associated tremor/ataxia syndrome, or FXTAS [61]. This is referred to as the “premutation state”, and predisposes to repeat expansion that may occur in the course of transmission of the affected chromosome from the mother to the child. FXTAS symptoms include tremor, ataxia, and cognitive decline. The disease mechanism is not well understood.

**GGGGCC repeat expansions in ALS and FTD**

Expansion of a GGGGCC repeat in a gene at the C9ORF72 locus on chromosome 9 is associated with more than 10% of familial cases and some sporadic cases of two different neurodegenerative diseases, amyotrophic lateral sclerosis (ALS), and frontotemporal dementia (FTD) [62–65]. This sequence motif conforms to the G4 consensus and has been shown to form quadruplexes as either RNA [66] or DNA [67]. In ALS, the death of motor neurons leads to stiff muscles and muscle wasting accompanied by difficulty in speaking, swallowing, breathing, and eventual death from respiratory failure. In FTD, neurons are lost from the frontal and/or temporal lobes, causing dementia that begins in late adulthood and is characterized by compulsive behaviors but—in contrast to Alzheimer’s disease—not by loss of memory. C9ORF72 GGGGCC repeat expansions have also been documented in other neurodegenerative diseases, including primary progressive aphasia [68] and depressive pseudodementia [69]. It is not understood why apparently identical molecular pathology gives rise to distinct symptoms.

Most individuals carry only two or three GGGGCC repeats at the C9ORF72 locus. Pathology is typically associated with at least 30 repeats, and in some cases, expansions may be much longer. Exact lengths can be difficult to determine, as GGGGCC repeats are not readily amplified by PCR. For diagnostic purposes, the standard assay is Southern blotting, which is more labor-intensive and requires more tissue, limiting the numbers of samples that can be analyzed and the accuracy of length measurements [70]. Like CGG repeat expansion diseases, GGGGCC repeat expansion exhibits anticipation, with age of presentation decreasing in successive affected generations [71]. However, in contrast to CGG expansions, GGGGCC expansions are variable in length among tissues of affected individuals. Thus, expansion must occur in somatic cells. In particular, long expansions of GGGGCC repeats have been reported in neural tissues, where they exert their pathological effects. Neural cells are non-dividing, so sequence expansion must occur in the absence of DNA replication.

**Quadruplex RNA: targets of pathology in G4 expansion diseases**

Most motifs that form G-quadruplex structures as DNA also do so as RNA, and the expansion-associated G4 motifs are not an exception. Cellular pathology in G4 expansion diseases can be ascribed in part to protein binding to quadruplexes in the pre-mRNA or mRNA transcript. G4 RNA in the FMR1 gene transcript titrates key factors that promote normal nuclear mRNA processing, contributing to some of the pathology resulting from CGG repeat expansions [72]. The GGGGCC motif forms parallel quadruplexes as RNA [66]. It is claimed that one source of pathology caused by GGGGCC expansion may be binding to RNA quadruplexes by the major nucleolar protein, nucleolin, which is claimed to deform nucleoli and disrupt normal nucleolar functions [67]. Altered nucleolar function would affect ribosome biogenesis and protein synthesis, and could cause many kinds of cellular pathology, though none specifically associated with ALS or FTD thus far.

**Are there other expandable G4 motifs?**

Currently, two additional unstable repeats are known that bear sequence motifs predictive of formation of G4 DNA within the gene itself and G4 RNA in the encoded transcript: the GGGGCT repeat in intron 1 of the NOP56 gene, on chromosome 20, associated with SCA36, an adult-onset disease of spinocerebellar ataxia and the CCCATGGTGGTTGGGACAG repeat in the PRNP gene, which encodes the prion protein.

Normal alleles of the NOP56 gene carry 5–14 GGGCC repeats, and expansions typically associated with disease are very long, in the range of 800–2,000 [73,74]. The expandable GGGGCT repeat is always found within a shared haplotype rather than occurring as a result of direct expansion. While relatively few patients have been analyzed, disease severity does not seem to vary with the length of the expansion, in contrast to GGGGCC and GGGGCC expansion diseases. The intronic position of the expandable GGGGCT repeat raises the possibility that G4 RNA contributes to pathology, as in CGG and GGGGCC repeat expansion diseases.

The normal prion gene contains 5 tandem G-rich repeats that encode a repetitive sequence within the larger prion protein, and are known as oligopeptide repeats (OPRs). Expansions of these repeats were one of the first genetic polymorphisms associated with CJD, a human transmissible spongiform encephalopathy [75,76]. CJD presents as dementia and results from destruction of neural cells in the brain due to aggregation by the prion protein. Expansions, typically from 5 to 10–14 OPRs, cause a dominant form of disease exhibiting early onset and slow progression. Those features led to initial identification of the disease caused by OPR expansion as Huntington’s disease phenocopy prior to definitive molecular disease and reclassification as familial CJD [77]. OPR expansion results in protein misfolding, evidenced by plaque formation by the prion protein in vivo and formation of insoluble protein aggregates in vivo [78,79]. It is an interesting possibility that G4 motifs are critical to disease-associated OPR expansions, but a role for them has not yet been demonstrated.

G-rich regions are notoriously difficult to amplify by PCR, so polymorphisms and instability at G4 motifs may go unscored by standard genomic analyses. The very recent identification of GGGGCC motif expansion as a cause of ALS and FTD [62–65] suggests that expandable G4 motifs may contribute to additional sporadic or familial neurodegenerative diseases.

**Quadruplex targets for antigenic variation in microbial pathogens**

Many pathogenic microbes evade the immune response by expressing a variety of different surface antigens. Switching surface antigen expression can be carried out by regulated recombination, which
moves a genetic cassette into an expression site. In the obligate pathogen Neisseria gonorrhoeae, which causes the venereal disease gonorrhea, antigenic variation depends on recombination. Variation is regulated by a recombination activator element that functions in cis to promote gene conversion of the active pilin expression locus, pilE, using a reservoir of silent pilS loci as sequence donors. The activator element is a G4 motif, (G3TG3TTG3TG3), and it must form a transcription-induced quadruplex structure to promote variation [80,81]. The conserved RecA recombination protein recognizes this specific quadruplex structure to stimulate strand exchange, and other G4 motifs do not support regulated recombination at the pilE locus [82].

G4-regulated antigen variation may also occur in other pathogens. A recent review [83] has compiled tantalizing examples of microbial pathogens which carry G4 motifs at sites of antigenic variation, which may occur by mechanisms related to those defined in N. gonorrhoea. Candidates include N. meningitides, which causes meningitis and other meningococcal diseases; the Borrelia spirochetes, among them B. burgdorferi, which causes Lyme disease; and Trypanosoma brucei, which causes sleeping sickness.

Quadruplex targets of transcriptional, translational, and epigenetic regulation

The sections below focus on recent genome-wide analyses that have identified quadruplexes as regulatory targets (summarized in Table 1). Numerous reports connecting specific G4 motifs with regulation of specific genes complement these genome-wide analyses, but the genome-wide studies are especially powerful because they provide new and unanticipated insights into how factors that act at G4 DNA or G4 RNA may contribute to cancer and genetic disease.

Quadruplex targets of transcriptional regulation

G4 DNA targets of XPD/XPB are enriched at the promoter

XPD and XPB are the helicase components of the general transcription complex TFIIH. They are essential for nucleotide excision repair; and specific mutant alleles are associated with three genetic diseases: Xeroderma pigmentosum, characterized by defective repair of damage from UV radiation, greatly increased predisposition to cancer, especially skin cancer and melanoma, trichothiodystrophy, characterized by distinctive developmental defects, especially brittle hair, but no cancer predisposition and Cockayne syndrome, characterized by developmental defects and progressive neurological degeneration, again without cancer predisposition [84,85]. We showed by biochemical analysis that XPD is a robust G4 helicase and that XPB binds but does not unwind G4 DNA, and by complementary ChIP-Seq mapping that XPB/XPD is greatly enriched at the TSS (Fig 3A), where 60% of binding sites overlap with G4 motifs (P = 0; [24]). Systems analysis shows that XPB/XPD is bound to G4 motifs in genes and pathways that are dysregulated in the diseases associated with deficiencies in these factors. Strikingly, highly significant enrichment of XPB/XPD was evident at G4 motifs near genes that escape the genome-wide shutdown of transcription that typically follows UV irradiation of human cells, and instead exhibit transcriptional upregulation in response to UV (P < 1e-12). That

Figure 3. Quadruplex targets of regulation.

(A) Three regions of human genes are G4hi ([G4]) and thus prone to quadruplex formation: upstream and downstream of the TSS and the 5' end of intron 1 (right). Quadruplexes upstream and downstream of the TSS are predominant among targets of XPB/XPD, while quadruplexes at the 5' end of intron 1 are predominant among targets of BLM. (B) Translation of mRNAs containing a quadruplex in the 5' UTR requires unwinding by the G4 RNA helicase, eIF4, which exposes the ribosome binding site to allow translation. (C) ATRX promotes epigenetic modifications at quadruplexes within G4hi telomeres and VNTRs.

© 2015 The Author EMBO reports Vol 16 No 8 | 2015

917
observation raises the intriguing possibility that not only failed UV repair but also dysregulated transcription may contribute to the predisposition to skin cancer and melanoma associated with XPB or XPD deficiencies.

XPB/XPD bind approximately 2% of all G4 motifs in the human genome. The features that distinguish XPB/XPD targets from other quadruplexes have yet to be defined. Quadruplex conformation is likely to be one of them. Transcriptional status and local DNA supercoiling (Fig 2) are also likely to be important, and interactions with other chromatin bound proteins may be necessary to promote stable associations.

**Quadruplex targets of BLM at the 5’ end of intron 1**

In order to determine whether BLM helicase might regulate transcription, we compared gene expression in fibroblasts from individuals with Bloom syndrome and normal individuals [86]. In BLM-deficient cells, mRNA expression of genes connected with cellular growth, proliferation, and cell death/survival was reduced. Altered expression of genes in these pathways may contribute to the primordial dwarfism that is a defining characteristic of Bloom syndrome. Genes downregulated in the absence of BLM were especially enriched in G4 motifs on the non-transcribed strand at the 5’ end of the first intron (FDR < 0.001), where approximately 20% of human genes carry a G4hi region that correlates with transcriptional pausing [87,88]. This suggests that BLM may normally unwind quadruplexes formed within the 5’ end of the first intron, thereby enabling transcription (Fig 3A).

**Quadruplex targets of translational regulation**

**G4 RNA targets of eIF4A-dependent translation**

Just as BLM helicase may smoothen the path for transcription through regions containing G4 DNA, the RNA helicase eIF4A may unwind G4 RNA to promote translation (Fig 3B). One well-documented function of this helicase is to scan the 5’ end of mRNA to reveal the site for translational initiation. Overexpression of eIF4A can drive development of T-cell acute lymphoblastic leukemia by inducing overexpression of other potent oncoproteins, such as MYC [89]. Ribosome footprinting showed that many genes that are translational regulatory targets of eIF4A bear a G4 motif in the 5’ UTR, and the (CGG)₄ motif appears to be a hallmark of mRNAs that depend on eIF4A for translation. These include oncogenes and transcription factors, classes of genes that were previously shown to be G4hi ([20]; see above). Dependence of these genes on eIF4A for efficient translation may provide an extra level of regulation that normally prevents overexpression of G4hi genes.

**Quadruplex targets of epigenetic regulation**

**ATRX**

The possibility that G4 motifs function as targets of epigenetic regulation was first raised by analyses of the SWI/SNF chromatin remodeler family member, ATRX [90]. ATRX deficiency is associated with a genetic disease, X-linked alpha thalassemia/mental retardation syndrome [90,91]. Somatic mutations in ATRX are common in pancreatic neuroendocrine tumors and tumors of the central nervous system, where they correlate with telomere maintenance by the telomerase-independent ALT pathway [92]. ATRX has robust G4 DNA binding activity, and it is enriched at telomeres, rDNA, and other sites. One of the targets of ATRX is a G-rich VNTR at the α-globin gene (CAGGGCGGGGG)n, which is predicted to form quadruplexes. In the absence of ATRX, this α-globin VNTR undergoes silencing, downregulating α-globin expression to cause the characteristic α-thalassemia. ATRX enables deposition of modified histones compatible with gene expression (Fig 3C). The G-rich α-globin VNTR reaches lengths of nearly 500 bp in some individuals, though it is much shorter in others, and the severity of α-thalassemia resulting from ATRX deficiency correlates with VNTR length [90].

**Perturbed replication promotes epigenetic instability at G4 motifs**

Epigenetic dysregulation is a well-documented feature of many tumors, and the mechanisms that cause this are not clearly understood. An intriguing series of experiments suggests that slowed replication at G4 motifs may be one source of epigenetic instability [93–95].

Epigenetic modifications at or near G4 motifs are altered in cells deficient in repair factors that promote smooth replication through quadruplexes, including FANCJ, BLM, or WRN helicases or the repair polymerase REV1 [93,94]. This causes epigenetic dysregulation evident either as increased expression of silenced genes or reduced expression of active genes. Conversely, an individual G4 motif can function as an epigenetic regulatory element that directs histone modifications at an adjacent promoter [95]. One model for how that may occur is as follows:

Histone deposition immediately after replication ensures correct propagation of specific histone marks. The normally rapid tempo of replication may be interrupted when the fork encounters a quadruplex structure, leading to diffusion of modified histones away from the stalled fork, and coating of the new daughter DNAs with histones bearing default marks that cannot exert appropriate regulation. This raises the possibility that some or all G4hi genes may be epigenetically labile, with chromatin structures tending toward the default setting with each round of replication. This would provide a level of control ensuring that G4hi oncogenes and transcription factors [20] could be expressed at high levels only if appropriate factors continued to be available following cell division.

**Do other disease-associated G4 helicases regulate gene expression?**

Further genome-wide analyses are necessary to map the quadruplex targets of other factors, an important step toward functional understanding of the G4 genome. Candidate regulators among the disease-associated human helicases include WRN and another RECQ family helicase, REQ4, deficient in Rothmund-Thomson syndrome, DNA2 and the XPD paralogs FANCJ, RTEL1, and CHL1 (CHL1/DDX11), deficient in Warsaw breakage syndrome, a cohesinopathy [96]. The iron–sulfur domain proteins (XPD, FANCJ, RTEL1, CHL1, and DNA2) are particularly interesting, as proteins in this family readily transport charge and they have been suggested to thereby relay intracellular signals [97].

G-quadruplex DNAs can themselves transport charge, over distances sufficiently long (100 nM) to be of potential utility for nano-devices [98]. This property could contribute a novel dimension to the functions of a quadruplex regulatory target in vivo. For example, iron–sulfur domain proteins may bind to intracellular
Sidebar B: In need of answers

- Which G4 motifs form quadruplexes in living human cells? A catalog of interactions of all endogenous quadruplex binding proteins with genomic DNA will provide a picture of this functional G4 genome.
- Which human G4 motifs are at risk of genomic instability? Do the same rules of G4 motif maintenance extend from yeast to human cells? If so, sites of instability in tumors are predicted to be enriched for G4L1 motifs.
- What is the mechanism of expansion of CGG motifs in germ cells or early embryonic development? How does it differ from the mechanism of expansion of GGGGCC motifs, which seems to occur in somatic cells?
- Does the prion gene G4 motif form a quadruplex structure in vivo, and does this promote sequence expansion in neurodegenerative disease?
- Which human G4 motifs are targets of transcriptional, translational, and epigenetic regulation?
- Do all the enzymes identified as G4 helicases also function as transcriptional regulators?

Detailed mechanistic understanding will enable rapid diagnosis of transcriptional, and translational regulation and chromatin structure.

Conclusions and future directions

We have begun to understand the connections between G-quadruplexes and human disease, at the levels of genomic stability, transcriptional, and translational regulation and chromatin structure. Detailed mechanistic understanding will enable rapid diagnosis of G4-associated disease, and development of new and better drugs with appropriate specificity. Recent progress now points to important questions for the immediate future (Sidebar B).

Acknowledgements

I thank the US National Institutes of Health and National Cancer Institute for supporting our research related to this review (PO1 CA77852).

Conflict of interest

The author declares that she has no conflict of interest.

References

G-embo reports


34. Paeschke K, Capra JA, Zakian VA (2011) DNA Replication through G-Quadruplex motifs is promoted by the Saccharomyces cerevisiae Pif1 DNA helicase. Cell 145: 678 – 691


41. Rossi ML, Ghosh AK, Bohr VA (2010) Roles of Werner syndrome protein in protection of genome integrity. DNA Repair (Amst) 9: 331 – 344


60. Mirkin EV, Mirkin SM (2014) To switch or not to switch: at the origin of repeat expansion disease. Mol Cell 53: 1 – 3


