



HAL
open science

Characterizing G-quadruplex formation : development and integration of novel assays

Yu Luo

► **To cite this version:**

Yu Luo. Characterizing G-quadruplex formation : development and integration of novel assays. Analytical chemistry. Université Paris-Saclay, 2023. English. NNT : 2023UPASF011 . tel-04953287

HAL Id: tel-04953287

<https://theses.hal.science/tel-04953287v1>

Submitted on 18 Feb 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Characterizing G-quadruplex formation: development and integration of novel assays

*Caractérisation de la formation de G-quadruplex :
développement et intégration de nouvelles méthodes*

Thèse de doctorat de l'université Paris-Saclay

École doctorale n°571 : Sciences chimiques : molécules, matériaux, instrumentation
et biosystèmes (2MIB)
Spécialité de doctorat: Chimie
Graduate School : Chimie. Référent : Faculté des sciences d'Orsay

Thèse préparée dans les unités de recherche **Chimie et modélisation pour la
biologie du cancer** (Université Paris-Saclay, CNRS, INSERM) et **Laboratoire
d'optique et biosciences** (Institut Polytechnique de Paris, CNRS, INSERM),
sous la direction de **Jean-Louis MERGNY**, directeur de recherche,
la co-direction de **Daniela VERGA**, chargée de recherche

Thèse soutenue à Paris-Saclay, le 17 Février 2023, par

Yu LUO

THESE DE DOCTORAT

NNT : 2023UPASF011

Composition du Jury

Membres du jury avec voix délibérative

Rachel MEALLET-RENAULT Professeur, Université Paris-Saclay (UMR 8214)	Présidente
Claudia SISSI Professeur, Università degli Studi di Padova, Dipartimento di Scienze del Farmaco	Rapporteur & Examinatrice
David MONCHAUD Directeur de recherche, Université de Bourgogne (UMR 6302)	Rapporteur & Examineur
Patrizia ALBERTI Chargée d'enseignement, Muséum National d'Histoire Naturelle (UMR 7196 - U 1154)	Examinatrice
Laurent LACROIX Chargé de recherche, HDR, Ecole Normale Supérieure (UMR 8197 - U 1024)	Examineur

Titre : Caractérisation de la formation de G-quadruplex : développement et intégration de nouvelles méthodes

Mots clés : Structure des acides nucléiques, G-quadruplexes, FRET, Caractérisation biophysique.

Résumé : Les séquences d'acides nucléiques riches en guanines peuvent adopter des structures secondaires non canoniques nommées G-quadruplexes (G4s). Ces structures ont été identifiées dans différents génomes et participent à de nombreux processus physiologiques. De nombreuses séquences G4s putatives (PQS) ont été identifiées par des analyses bioinformatiques et séquençage par immuno-précipitation de la chromatine (ChIP-seq). Par contre, trouver une séquence relativement riche en guanine ne signifie pas forcément qu'elle forme une structure G4. Le but de cette thèse se concentre sur le développement de nouvelles méthodes d'étude *in vitro* pour la caractérisation de G4s à haut débit, et la validation d'un grand nombre de PQS.

Le transfert d'énergie par résonance de type Förster (FRET) est un puissant outil pour le suivi de la dénaturation d'un G4 intramoléculaire fonctionnalisé aux extrémités avec deux sondes: un chromophore donneur (FAM) et un accepteur (TAMRA). À faible température, les deux extrémités sont à proximité, permettant la désactivation de la fluorescence du donneur par FRET. Quand la température augmente, le G4 est dénaturé et les deux extrémités s'éloignent, ce qui restaure graduellement la fluorescence du donneur. La température de dénaturation (T_m) peut être obtenue à partir de la courbe de FRET-melting.

Une nouvelle méthodologie nommée FRET-melting compétitive (FRET-MC) a été développée, permettant de suivre 48 échantillons en duplicata en 2 heures. Le FRET-MC se base sur la compétition d'un ligand sélectif pour les G4s (PhenDC3) entre un G4 reporteur (FAM-Tel21-TAMRA, F21T) et un compétiteur en large excès dont la structure est inconnue. L'interaction du PhenDC3 avec le G4 formé par F21T induit une stabilisation mesurée par l'augmentation du T_m d'environ 23°C. L'excès de compétiteur G4 va piéger le PhenDC3, ramenant le T_m de F21T à la valeur obtenue sans ligand. Une séquence compétitrice qui ne forme pas une structure G4 n'aura pas d'impact sur l'interaction du PhenDC3 avec F21T. Le FRET-MC est une méthode tolérante une

grande variabilité de compétiteurs, excepté ceux présentant une faible stabilité thermique, qui se retrouvent sous forme monocaténaire à la température où F21T commence à se dénaturer.

Le FRET isotherme (isso-FRET) a été développé comme une alternative compatible avec des G4s peu stables thermiquement. L'iso-FRET exploite deux RNAs : le 37Q (37- quencher) formant un G4 lorsque stabilisé par le PhenDC3, et le F22 (FAM-22) partiellement complémentaire à 37Q. À ce système est ajoutée une séquence compétitrice non marquée. Si le compétiteur ne forme pas de G4, le PhenDC3 stabilise 37Q, empêchant son interaction avec F22, ce qui permet un fort signal de fluorescence de F22. Au contraire, un excès de compétiteur formant un G4 piégera le PhenDC3 qui ne sera plus disponible pour stabiliser 37Q. Il en résulte l'hybridation de 37Q avec F22, provoquant la désactivation de la fluorescence de F22. L'iso-FRET peut être utilisé à température 25°C ou physiologique (37°C), ce qui le rend compatible avec les compétiteurs G4s possédant une faible stabilité thermique.

En parallèle, la conformation des séquences riches en CG formant des G4s a été réalisée. La structure des séquences riches en GC est étroitement liée au nombre de cytosines, mais également au ratio de $[K^+] / [Na^+]$ présent dans le tampon. Les résultats spectroscopiques montrent que CEB25 tolère des boucles continues de cytosines dans un tampon mixte potassium / sodium. Cela implique que les séquences G4s riches en cytosine peuvent toujours adopter une structure G4 dans l'environnement intracellulaire. Les résultats d'UV-melting mettent en évidence que la présence de cytosines dans les boucles G4 ne diminuent pas leur stabilité, mais augmente la probabilité de former une tige-boucle ou un ADN palindromique bicaténaire en compétition. L'emplacement des PQS riches en cytosines a été étudié par analyse bioinformatique et indique une localisation au niveau des pré-ARNm.

Title : Characterizing G-quadruplex formation: development and integration of novel assays

Key Words : Nucleic acids structures, G-quadruplexes, FRET, Biophysical characterizations.

Abstract : Guanine-rich nucleic acid sequences can generate four-stranded, noncanonical secondary structures called G-quadruplexes (G4). G4 structures have been identified in the genomes of different species and are involved in physiological processes. Numerous putative G4 sequences (PQS) have been mined by G4 predicting algorithms and G4 chromatin immunoprecipitation sequencing (ChIP-seq). However, finding a sequence relatively rich in guanines does not necessarily mean it can form a G4 structure. The main aim of this PhD thesis is to develop novel *in vitro* high-throughput G4 characterization assays, suitable for validating a vast amount of PQS identified by *in silico* G4 prediction methods and ChIP-seq.

Förster Resonance Energy Transfer (FRET) is a powerful tool in characterizing intramolecular G4 structures that were dual-labeled by a chromophore (*i.e.* FAM) and a corresponding quencher (*i.e.* Tamra): at the low temperature, the two ends of the G4 are in close proximity and the FAM fluorescence is quenched *via* a FRET mechanism; with the temperature increasing, G4 unfolds and two ends are split apart, restoring gradually FAM fluorescence. Melting temperature (T_m) of a G4 structure could be calculated based on the FRET-melting curve.

In this context, a new methodology called the FRET-melting competition (FRET-MC) assay has been developed, which allows to follow 48 duplicated samples within 2 hours. FRET-MC is based on the competitive binding of a selective G4 ligand (PhenDC3) between a labeled G4 reporter (FAM-Tel21-TAMRA, F21T) and an unknown competitor present in a large excess. PhenDC3 stabilizes the G4 structure of F21T and leads to an increase of the T_m of F21T of about 23 °C. An excess of G4 competitors can trap PhenDC3, leading to a decrease of the T_m of F21T back to the T_m value of F21T alone. In contrast non-G4 competitors have little influence on the F21T-PhenDC3 interaction. FRET-MC works well in most cases, but cannot be used to pick up G4s with low thermal stability, as these weak G4 behave as ssDNA at the temperature where F21T starts to melt.

For this reason, an alternative isothermal FRET assay compatible with weakly stable G4s was developed. Iso-FRET exploits two labeled RNA probe strands: one of them 37Q (37-quencher) form a G4 thanks to the binding of PhenDC3, and the second one F22 (FAM-22) is partially complementary to 37Q. To this system an unlabeled competitor sequence is added. When the competitor is not a G4, PhenDC3 remains bound to 37Q and stabilizes its G4 structure, preventing duplex formation between 37Q and F22, allowing a high fluorescence signal of F22. On the contrary, an excess of a G4 competitor traps PhenDC3, which is no longer available to bind to 37Q, which in turns allows this oligonucleotide to hybridize to F22 resulting in fluorescence quenching. Iso-FRET can be performed at 25 °C and 37 °C (physiological temperature), which are acceptable to thermolabile G4 competitors.

In parallel, a study focusing on the conformation of G4-prone GC-rich sequences which are rich both in cytosines and guanines was conducted. The structures of GC-rich sequences were dependent on cytosine content and $[K^+] / [Na^+]$ ratio in buffer. Spectroscopic results shown that the G4 structure (CEB25) tolerates two three continuous cytosines tracks in potassium / sodium mixed buffers containing 40 mM or higher KCl concentration, implying that cytosine contained G4-forming sequences can still adopt a G4 structure in the intracellular environment. UV-melting results evidenced that the presence of cytosines in G4 loops does not decrease G4s stability, but rather increases the probability of forming a competing structure, either a hairpin or a duplex. As bioinformatics analysis indicates that the majority of cytosine-contained PQS located at pre-mRNAs, it will be interesting to transpose our experiments on DNA sequences to RNA motifs, in order to determine if the latter is more or less prone to hairpin / G4 competition.

Content

Acknowledgments.....	3
List of abbreviations.....	5
Chapter I. Introduction.....	7
1. Secondary structures of nucleic acids.....	7
1.1 Canonical nucleic acids structures.....	8
1.2 Four-stranded nucleic acids structures.....	12
2. G-quadruplexes in genomes and their biological significance	20
2.1 Mapping G-quadruplexes in human genomes and other species	21
2.2 G-quadruplexes in gene promoters.....	22
2.3 G-quadruplexes in gene expression: beyond downregulation.....	24
2.4 G-loop: G-quadruplexes co-exist with R-loops.....	25
2.5 G-quadruplexes in gene replication	26
3. Characterization of G-quadruplexes <i>in vitro</i>	27
3.1 UV-visible absorbance spectra.....	28
3.2 Circular Dichroism (CD) spectrum	32
3.3 Nuclear magnetic resonance (NMR).....	35
3.4 Gel electrophoresis.....	36
3.5 Intrinsic fluorescence	37
3.6 Fluorescence “light-up” assays	38
3.7 Additional methods of interest.....	39
4. Specific G-quadruplex ligands.....	39
4.1 Small molecular G-quadruplexes ligands.....	40

4.2 G4 interacting proteins	45
5. Thesis objectives	47
Chapter II. FRET melting competition assay	51
Chapter III. Isothermal FRET competition assay	85
Chapter IV. G-quadruplexes characterization <i>in vitro</i>	121
Chapter V. A sodium / potassium switch for G4-prone GC-rich sequences.....	123
Chapter VI. General conclusion and perspectives	157
Appendix (articles not included in the main thesis manuscript)	161
References.....	217
Abstract in English	235
Resumé en français.....	239

Acknowledgments

First of all, I would like to thank Dr. Florence MAHUTEAU-BETZER, Dr. Marie-Paule TEULADE-FICHOU and Dr. François HACHE for allowing me to complete my thesis in their units. I am also grateful for their interesting remarks and good advises.

I would like to appreciate my thesis director Dr. Jean-Louis MERGNY and Dr. Daniela VERGA. I am very happy that my research was supervised by such nice, kind, delicate and supportive people, with outstanding scientific view and interesting lifestyle. I am truly grateful for all things they did for me, including not only methods and techniques connected to my research, but also for helping me to become more logical, critical, consistent and organized.

I am grateful to Dr. Laurent LACROIX for being my thesis monitoring committee and jury composition, for his meaningful and helpful advising about my research and career. I thank Pr. Claudia SISSI and Dr. David MONCHAUD for being my thesis reporters, for their critical examination and helpful suggestions about my thesis manuscript. I also thank Pr. Rachel MEALLET-RENAULT and Dr. Patrizia ALBERTI, for their work and time on my thesis and defense.

I would like to thank Dr. Anton GRANZHAN and Dr. Lionel GUITTAT for their openness to discussions and assistance for experiments and data dealing. Special thanks to Anne CUCCHIARINI and Dr. Zackie AKTARY for their support on experiments and all joy they bring to the laboratory.

I would like to thank Dr. Samir AMRANE for welcoming me to his laboratory at Institut Européen de Chimie et Biologie (IECB, Pessac, France). I was given an exciting opportunity to learn principles deeply and perform nuclear magnetic resonance spectrometry experiments. I would like to thank Dr. Julien MARQUEVIELLE and Dr. Gilmar SALGADO for their assistance and warm reception. I would like to thank all people that helped me with my research. I thank Dr. Martina Lenarčič Živković and Dr. Lukáš Trantírek in Central European Institute of Technology (CEITC, Brno, Czech Republic), for their assistance at NMR spectra in living cells.

I would like to thank all members of our friendly teams. Thank our secretaries, Nathalie (CMBC, Inst. Curie) and Laure (LOB, Polytechnique) for their administrative support. I thank Charles-Henri, Corinne, Joseph, Rahima, Jean and Liliane (CMBC, Inst. Curie); and Roxane, Jiawei, Auriane, Bahar, Rongxin, Júlia, Sophie, Kilolo, Rivo (LOB, Polytechnique) for their assistance and nice company. I also keeping warm memories of days shared with all former members, including Thibaut, Jaime, Oksana, Eugénie, Chloé in Inst. Curie; and Dale, Anastasia, Seongbin in Polytechnique.

I would appreciate all relevant affiliations, for their financial support in the personal emolument (China Scholarship Council 201906340018), experimental sites and equipment, and research fundings.



I would like to keep the last grateful moment for my family members, particularly my brother, for unlimited love, endless support and eternal believe in his young sister.

List of abbreviations

Ribonucleic acid	RNA
Deoxyribonucleic acid	DNA
Adenine	A
Guanine	G
Cytosine	C
Thymine	T
Uracil	U
Double-stranded	<i>ds</i>
Single-stranded	<i>ss</i>
Hydrogen bonds	H-bonds
B-type duplex DNA	B-DNA
Base pairs	bp
A-type duplex DNA	A-DNA
Left-handed DNA	L-DNA
Messenger RNA	mRNA
Transfer RNA	tRNA
Human immunodeficiency virus	HIV
Untranslated region	UTR
RNA interference	RNAi
Micro RNA	miRNA
Small interfering RNA	siRNA
G-quadruplexes	G4
Polyethylene glycol	PEG
Melting temperature	T_m
Nuclear magnetic resonance	NMR
Circular dichroism	CD
Adeno-associated virus	AAV
Putative G-quadruplex sequences	PQS
Non-coding RNAs	ncRNAs
Ribosomal RNA	rRNA
Chromatin immunoprecipitation sequencing	ChIP-seq
Transcription start site	TSS
Tumor suppressor genes	TSGs
Single-nucleotide variants	SNVs

Human telomerase reverse transcriptase	hTERT
Nuclease hypersensitive element	NHE
Nuclease hypersensitive polypurine-polypyrimidine element	NHPPE
Nucleolin	NCL
G4-binding proteins	G4BPs
Long terminal repeat	LTR
Human immunodeficiency virus-1	HIV-1
Myc-associated zinc-finger protein	MAZ
Cross-linking immunoprecipitation sequencing	CLIP-seq
RNA-protein binding sites	RBPs
Transcription-replication conflicts	TRCs
Origin recognition complex	ORC
Glycosidic bond angles	GBA
Excitation wavelength	λ_{ex}
Emission wavelength	λ_{em}
Reversed-phase high-performance liquid chromatography	RP-HPLC
Size exclusion chromatography	SEC
Analytical ultracentrifugation	AUC
Mass spectrometry	MS
Dimethyl sulfate	DMS
N-methyl mesoporphyrin IX	NMM
Acridine orange	AO
Topoisomerase II	Topo II
RNA polymerase I	Pol I
Human protection of telomeres 1	POT1
Copper-catalyzed azide-alkyne cycloaddition	CuAAC
Metal-complex strain-promoted alkyne-azide cycloaddition	SPAAC
Restricted intramolecular rotation	RIR
Thioflavin T	ThT
Adenosine triphosphate	ATP
Telomerase RNA component	TERC
Förster resonance energy transfer	FRET
FRET-melting competition assay	FRET-MC
Human telomerase RNA gene	<i>hTERC</i>

Chapter I. Introduction

1. Secondary structures of nucleic acids

Nucleic acids are biological macromolecules composed of nucleotides which constitute the genetic material for all known life forms. As the basic units of natural nucleic acids, nucleotides are composed of a nucleobase (also known as a nitrogenous base), a five-carbon sugar (ribose or deoxyribose), and a phosphate group. As shown in **Figure 1a**, depending on the five-carbon sugar (ribose, in which a hydroxyl group is present on the 2' carbon atom ($R = OH$) or deoxyribose, in which a hydrogen atom is present on 2' carbon atom ($R = H$)), natural nucleic acids are divided into ribonucleic acid (RNA) and deoxyribonucleic acid (DNA). The 2'H / OH difference has a huge influence on the sugar conformation: in general, the ribose of RNA adopts a C3'-*endo* conformation, whereas the deoxyribose of DNA can adopt both C2'-*endo* and C3'-*endo* conformations [1] (**Figure 1b**), and this sugar pucker preference will affect the ability of a purine to switch to a *syn* conformation (see below).

Nucleic acids are constituted by nucleotides containing natural different nucleotides: adenine (A), guanine (G), cytosine (C) and thymine (T) in DNA or uracil (U) in RNA. Differently, Adenosine, Guanosine, Cytidine, and Thymidine or Uridine correspond to the *nucleoside* (**Figure 1c**). The phosphate group attached to the 5' carbon atom and the hydroxyl group on adjacent 3' carbon atom form phosphodiester bonds and link nucleotides together within a single-stranded nucleic acid chain. Aside from the five natural nucleotides, a number of natural and man-made nucleotides including modified bases, sugar rings, and phosphate group analogs have been studied, and their properties are now better understood [2].

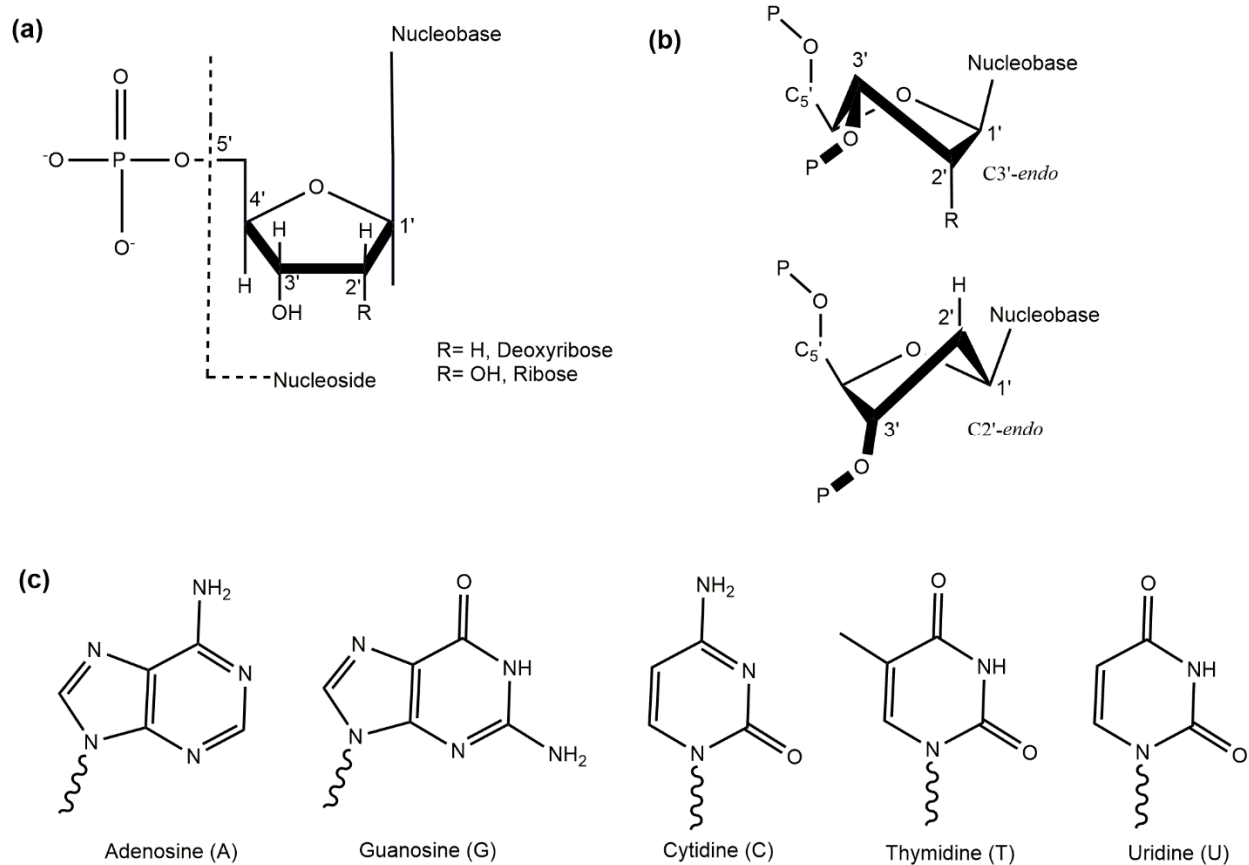


Figure 1 (a) Schematic representation of the nucleotide structure, which includes a nucleobase, a five-carbon sugar, and a phosphate group. (b) Sugar puckering: conformations of C3'-endo and C2'-endo. (c) Chemical structures of nucleobases, ~ corresponds to the bond to the C1' carbon atom of the sugar.

1.1 Canonical nucleic acids structures

1.1.1 Double-stranded DNA

In 1953, Watson and Crick determined the helical structure of double-stranded DNA (*ds*), based on X-ray diffraction. The structure consisted of two complementary single-stranded (*ss*) sequences held together by nucleobase stacking and the formation of inter-strand hydrogen bonds connecting the nucleobases. These hydrogen bonds (H-bonds) are now often referred to as 'Watson-Crick hydrogen bonds'. There are two H-bonds in an A-T base pair (bp), and three in a C-G bp [3, 4] (**Figure 2a**).

The vast majority of DNA probably adopts a B-type duplex structure (B-DNA, **Figure 2b**) under physiological conditions. Canonical B-DNA consists in an antiparallel right-handed double-helix constituted by two complementary strands, where the two antiparallel stands (5'-3' / 3'-5') are held together by hydrogen bonds in A-T and C-G base pairs. This double-helix has a diameter of about 2 nm and a persistence length of up to 53 nm on mica surface [5]. The double-helical structure presents two different grooves, named the major groove and the minor groove, twisting around the central axis on opposite sides. The major groove is relatively shallow and wide, while the minor groove is deep and narrow [6]. A complete helical turn of B-DNA has a pitch of about 3.4 nm and contains approximately 10 bps.

When dehydration occurs *in vitro* (*i.e.*, when the relative humidity drops to 75% or less), the DNA duplex switches reversibly to an A-type duplex (A-DNA) [7, 8]. Compared to B-DNA, A-DNA is like a twine that has been tightened more. A-DNA reduces the pitch to 2.8 nm and contains 11 bp, and increases the diameter from 2.3 nm (for B-DNA) to 2.5 nm; the angle (twist) between adjacent base pairs changes from 36° to 33°. The sugar ring conformation changes from C2'-*endo* (for B-DNA) to C3'-*endo* for A-DNA [9]. Of note, the base pairs are no longer perpendicular to the helix axis.

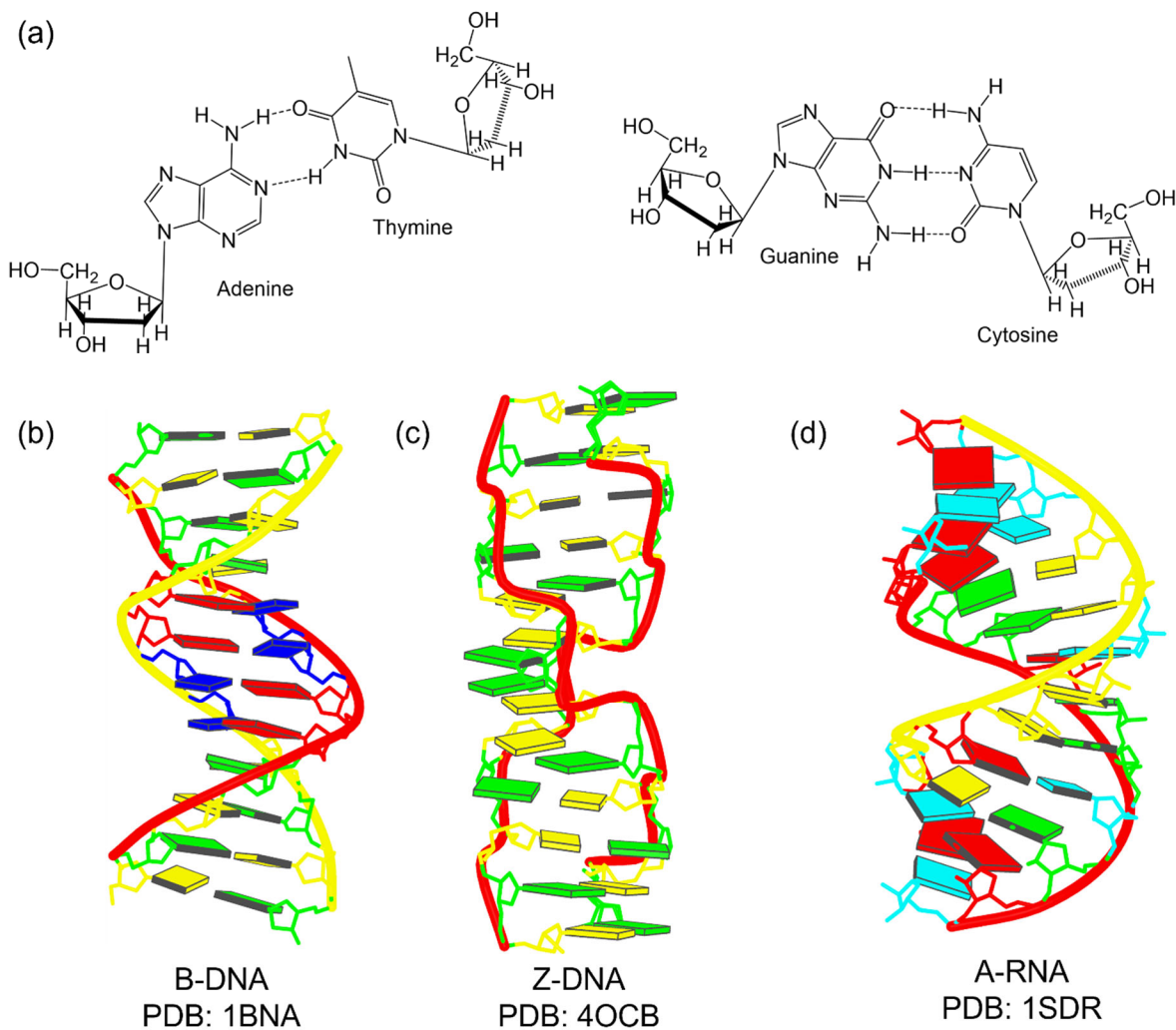


Figure 2 (a) Left: A and T are connected by two Watson-Crick hydrogen bonds. Right: C and G are connected by three Watson-Crick hydrogen bonds. (b) The first published B-DNA structure with the sequence $d(\text{CGCGCGCGCGCG})$ and its complementary strand. (c) Z-DNA of $d(\text{CGCGCGCGCGCG})_2$. (d) A-RNA of $r(\text{UAAGGAGGUGAU})$ and its complementary strand. Models were generated with Web 3DNA 2.0 [10], Color code for base rectangular blocks: Adenine, red; Cytosine, yellow; Guanine, green; Thymine, blue; Uracil, cyan.

Additionally, an unusual duplex configuration called Z-DNA (**Figure 2c**) may also form under specific conditions. Z-DNA is significantly different from the two configurations described above. Z-DNA is left-handed, and all phosphate groups in the sugar-phosphate chain are arranged in a zig-zag shape, giving the structure its "Z"-DNA name. It was first discovered by Wang *et al.* when they studied single crystal X-ray diffraction pattern of the $d(\text{CGCGCG})$ sequence [11]. Several chemical agents including spermine, hexammine cobalt(III), and ruthenium complexes can convert B-DNA into Z-DNA [12]. Z-DNA has been implicated in gene activation, chromatin remodeling and large-scale genomic deletions in mammalian cells

[13]. Z-DNA also has the potential to be a therapeutic target in cancer treatments [14] and was recently reported to be a major component of extracellular DNA in bacterial biofilms [15]. Z-RNA is also biologically relevant [16].

Besides A-, B- and Z-DNA, there are other types of DNA helices. For example, D-form is a type of right-handed twist composed of poly (dA - dT) and poly (dG - dC); each turn contains 8 bp [17]. In addition, non-natural nucleic acids may adopt different helical arrangements. For example, left-handed DNA (L-DNA) can adopt the exact mirror-image helix of the natural DNA conformer [18], sharing the same physical characteristics, including solubility, duplex stability and selectivity, but with the opposite chirality. L-DNA has been used to develop a universal microarray platform capable of simultaneously analyzing multiple molecular parameters [19] and mirror-image aptamers called *spiegelmers*.

A few years after the discovery of the classical antiparallel B-DNA structure, the first parallel (5'-3' / 5'-3') double-stranded DNA helix was reported in 1961. X-ray diffraction illustrated the compact parallel double helix formed by poly (A) under acidic conditions, in which adenine residues in two opposite parallel helical strands form three hydrogen bonds [20]. Two d(A₁₅) strands adopted a right-handed parallel-stranded double helix in sodium (pH = 3) with high thermal stability; adenines in this duplex were held together by AH⁺-H⁺A base pairs; while d(A₁₅) existed as a structured single helix under neutral conditions (pH = 7) [21].

1.1.2 Single- and double-stranded RNA

In contrast to genomic DNA, RNA is generally not present as a perfectly-matched duplex in eukaryotic cells. This does not mean that common RNAs such as messenger RNA (mRNA) and transfer RNA (tRNA) are single-stranded, as they adopt complex secondary and ternary folds [22]. Additionally, nature contains numerous positive single-stranded RNA viruses, including Hepeviridae, Coronaviridae and Arteriviridae [23]. These single-stranded RNA viruses may often be pathogenic. Human immunodeficiency virus (HIV) codes for a reverse transcriptase enzyme, which allows single-stranded RNA to be reverse-transcribed into DNA. The new SARS-CoV-2 virus is a member of the Coronaviridae family. Surprisingly, its RNA has been claimed to be able to integrate into the genome of cultured human cells, resulting in expression in patient-derived tissues [24]. It is not unusual to find *dsRNA* in a natural context. *dsRNA* comprising in the 5' untranslated region (5'-UTR) of a mRNA can effectively inhibit translation. Chang *et al.* mapped *dsRNA* in HeLa and human embryonic kidney (HEK 293T) cells in 2016. *dsRNA* can form across long distances and many of them are long (> 200 nt), and pervasively occur with alternative structures where one sequence can base pair with two or more different partners [25]. Existence of *dsRNA* also provides the basis of RNA interference (RNAi), which is a naturally occurring gene silencing mechanism. Two different major types of small RNAs,

microRNA (miRNA) and small interfering RNA (siRNA), may downregulate gene expression by targeting to specific mRNAs [26].

As with DNA duplexes, RNA helices are composed of two complementary strands that are oriented antiparallel to each another. These two strands are connected *via* Watson-Crick hydrogen bonds between nucleobases. The hydroxyl group on the 2' carbon atom prevents the ribose from adopting the C2'-*endo* conformation, allowing only the C3'-*endo* form. This ribose conformation eliminates the possibility of forming a B-type RNA double-helix, making A-type the regular type of RNA duplex (**Figure 2c**). Similar to A-DNA, A-RNA has a relative shorter pitch (3.1 nm) than B-DNA [27]. In comparison to DNA-DNA and DNA-RNA hybrid complexes, RNA-RNA duplexes tend to be more stable [28, 29], possibly as a result of the rigidity of RNA strands [29] and the effects of hydration [30].

1.2 Four-stranded nucleic acids structures

In addition to the double-helices described above, DNA and RNA can adopt alternative secondary structures involving more than two strands [31, 32]. In particular, one can form three-stranded structures or triplexes (based on base triplets such as GC•C+, TA•T and CG•G, *etc.*) [33, 34] and four-stranded structures such as G-quadruplexes, *i*-motif and other oddities [35, 36]. Penta-stranded structures have also been described with non-natural bases such as isoguanine [37].

1.2.1 G-quadruplex

G-quadruplexes (G4) constitute a family of secondary structures generated by G-rich nucleic acid sequences (generally involving 1, 2 or 4 strands). Four Guanine bases are connected by Hoogsteen hydrogen bonds (**Figure 3a**) to form a *G-tetrad*, or *G-quartet*, the basic unit to form G4, which had been at first proposed by Gellert *et al.* in 1962 [38]. The π -stacking of two or more G-tetrads generates a quadruplex structure (**Figure 3b**). In 1989, Williamson *et al.* observed that telomeric G-rich sequences may adopt a quadruplex fold, as shown by anomalous electrophoretic mobility in polyacrylamide gels [39]. Since then, the interest for G4 ushered in explosive growth.

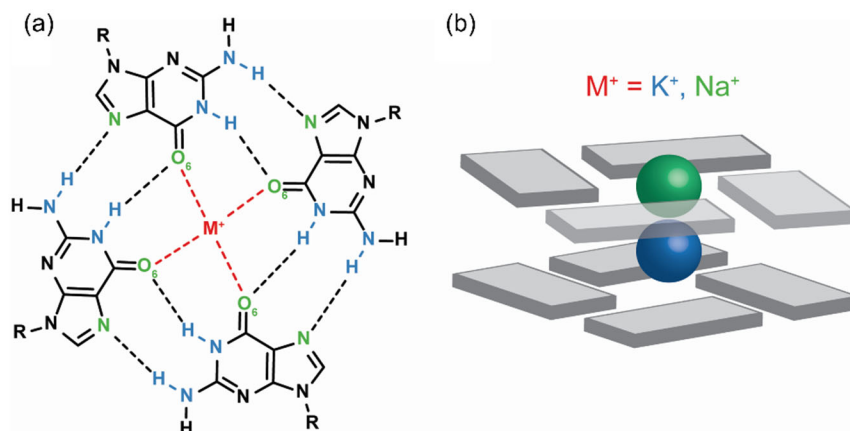


Figure 3 (a) Guanine tetrads are composed of four guanines that are linked together by eight Hoogsteen hydrogen bonds (donor and acceptor groups in blue and green, respectively). A metal cation is selectively coordinated by guanine O6 (red). (b) Two G-tetrads stack to form a G-quadruplex. Cations with larger ionic radii are located between tetrads (case of K^+ ; blue), while smaller ones may also coordinate within the plane of the tetrads, or adopt an intermediate position (case of Na^+ ; green). Figures are adopted from [40].

G-quadruplexes are highly polymorphic structures. As guanine nucleotides may adopt *anti* and *syn* conformations (**Figure 4a**), the arrangement and combination of *anti* and *syn* guanines result in different types of G-tetrads with varied grooves (**Figure 4b**) and stacking features. Biophysical parameters of the classical duplexes and G-quadruplexes [41] are summarized in **Table 1**. Furthermore, G4 can also be classified into parallel, antiparallel, and hybrid structures according to *anti* / *syn* conformations of the guanines [41] (**Figure 5**). Based on the molecularity, G4 can be categorized into monomolecular (intramolecular), bimolecular or tetramolecular (intermolecular) or higher-order structures. According to loop types, anti-parallel structures can be further subdivided into chair and basket configurations.

Differing by the group on the 2' carbon atom of the sugar ring, G4 conformations of DNA and RNA are substantially diverse. RNA G4 tend to be more stable. The 2'-hydroxyl group in the ribose sugar within RNA G4 grooves contributes to attract water molecules, leading to a more stable structure than the one generated by DNA G4 grooves [42]. The 2'-hydroxyl imposes additional steric constraints on the G4 topology, where it prevents the base from being oriented in the *syn*-conformation, strongly favoring the *anti*-conformation *via* constraints on the glycosidic torsion angle. As show in § 1.1.2 *Single- and double-stranded RNA*, the 2'-hydroxyl imposes constraints on sugar pucker; it only allows the C3'-*endo* form. As a result, RNA G4 structures are nearly always restricted to the parallel topology, in which all four strands go in the same direction [43]. Compared to RNA G4, the orientations of G-strands in DNA G4s may vary. DNA G4s can adopt parallel, antiparallel, and hybrid structures.

Table 1. Biophysical parameters of DNA duplexes and DNA G-quadruplex

Structural type	B-DNA	A-DNA	Z-DNA	G-quadruplex
Rise per base pair (Å)	3.4	2.9	3.7	3.3
No. of bases per turn (bp)	10.5	11.0	12.0	12.0
Twist (°)	36.7	32.7	-10/-50	30
Groove width (Å)	11.7/5.7	2.7/11	8.5	12.0/14.5/18.0*
Strand polarity**	ap	ap	ap	variable

* G4 structures may contain three types of grooves: narrow, middle and wide.

** "ap" stands for antiparallel: the two strands run in opposite orientations.

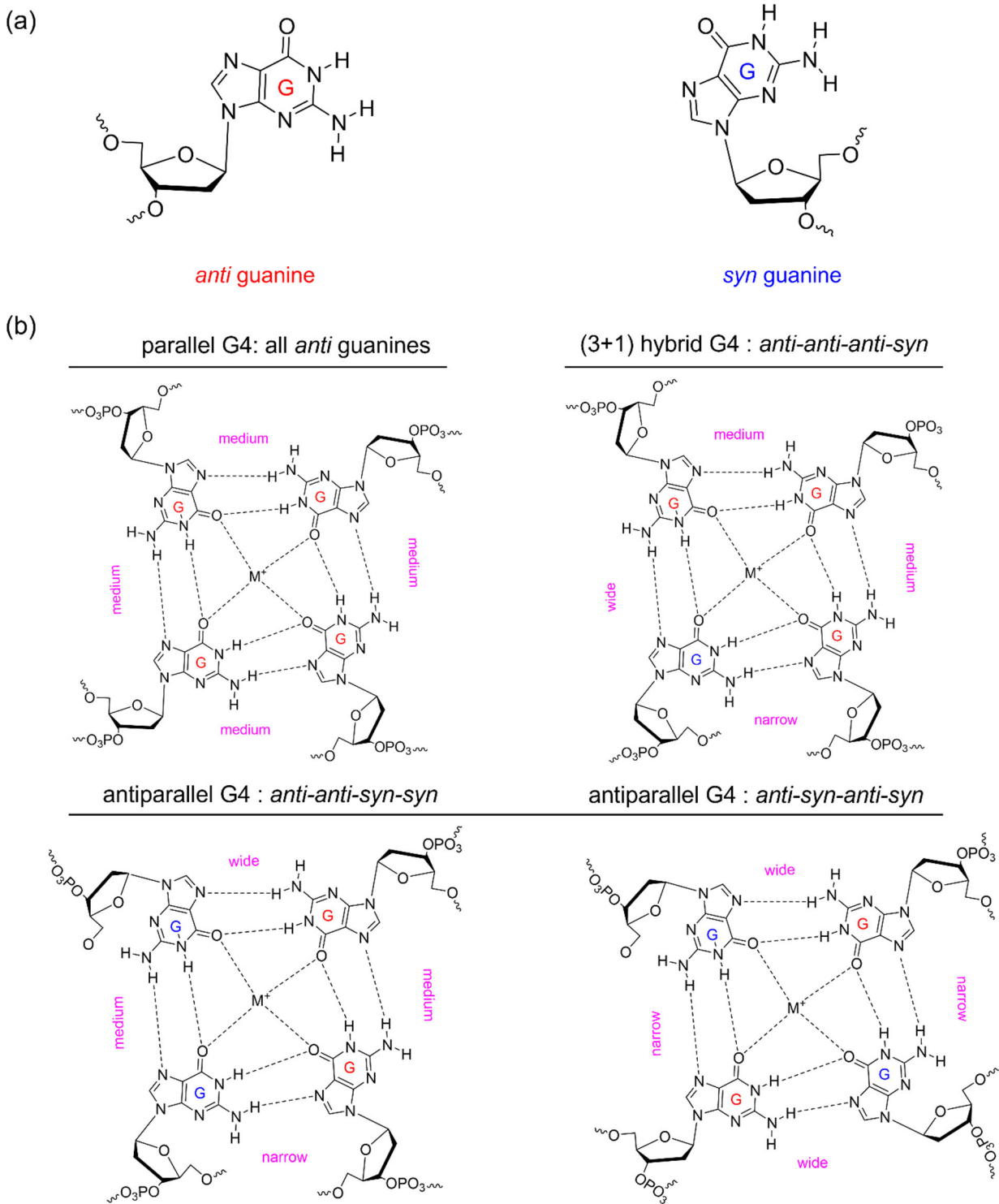


Figure 4 (a) *anti* and *syn* conformations of guanines. (b) different G-tetrads with varied grooves in polymorphic G4 conformations.

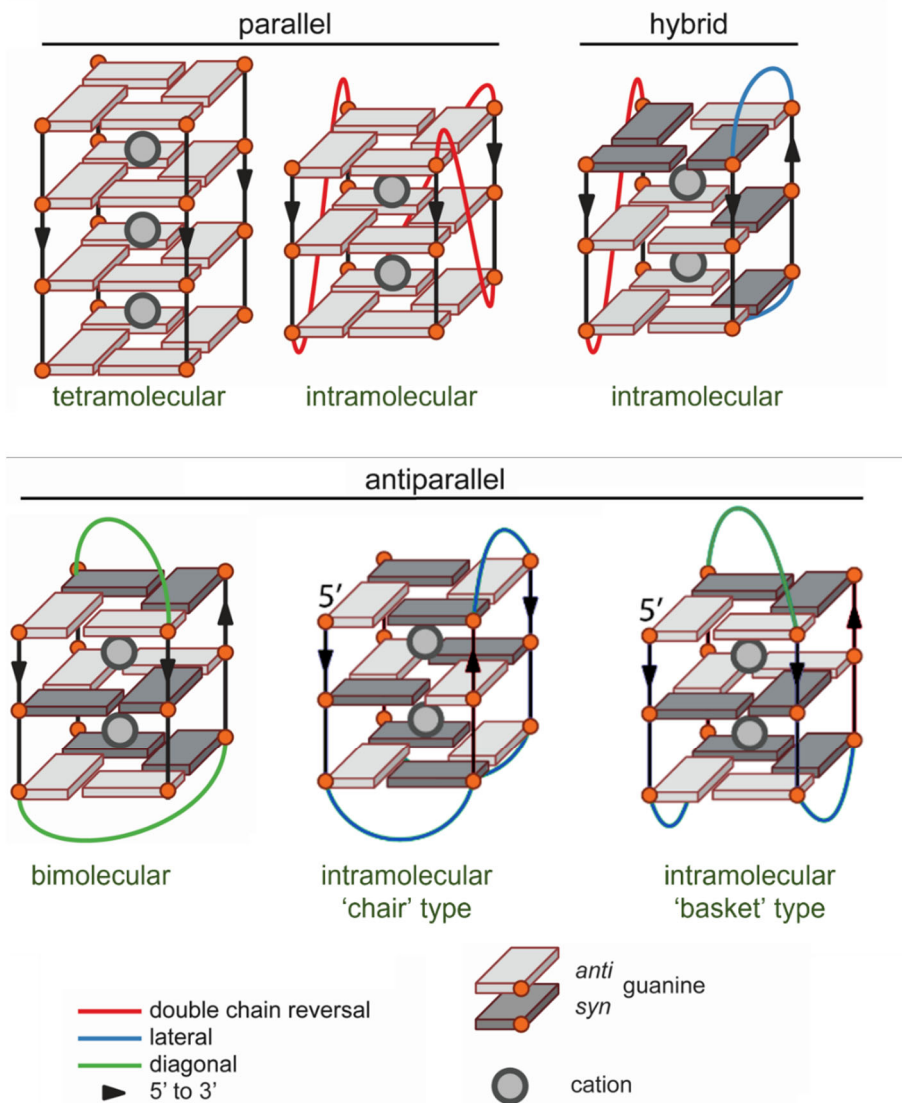


Figure 5 G-quadruplexes can fold into a variety of topologies that differ mainly by the relative orientation and number of strands (1 to 4), the number of tetrads (at least 2), and the geometry of the loops. The figure is adapted from [40].

Topologies and stabilities of G4 are affected by several factors:

- i)** Temperature: All nucleic acid secondary structures are enthalpy-driven, G4 is no exception: it tends to unfold at high temperature.
- ii)** It is possible to stabilize G4 structures with specific small molecules [44, 45].
- iii)** Ionic environment: Metallic cations can stabilize G4 by laying between two G-tetrads (sodium and potassium) and coordinating with eight 6' O atoms. Thanks to their small ionic radius, sodium can also fit in the core of G-tetrads and form a tetra-coordinated complex. The general trend of G4 stabilized by monovalent and divalent cations is $\text{Sr}^{2+} > \text{Ba}^{2+} > \text{K}^+ > \text{Ca}^{2+} > \text{Na}^+, \text{NH}_4^+, \text{Rb}^+ > \text{Mg}^{2+} > \text{Li}^+ \geq \text{Cs}^+$ [46].

iv) Molecular environment: Crowding conditions have been reported to preferentially stabilize G4 structures. Polyethylene glycol (PEG) is a widely used crowding agent, which may convert G4 topologies from anti-parallel to parallel in sodium buffers [47]. However, other crowding agents may have different effects – PEG is not a true crowding agent, and stabilization may not only be related to crowding effects as observed for example for i-DNA [48]. Regarding G-quadruplexes, PEG binds G4 based on conformational selection leading to the observed conformational transition [49].

v) Loop length: In potassium buffers, the total loop length is inversely correlated to G4 thermal stability: each added base leads to a 2 °C drop in T_m [50], and only a few G4s with a loop of more than 7 nt have been reported (*i.e.*, [51]). Loop length also effects G4 topology. Generally, very short linkers (1 or, to a lesser extent, 2-nt long loops) impose chain reversal loops, which impose a parallel topology. On the other hand, longer linkers are compatible with all types of loops and non-parallel topologies may be observed. A loop of moderate length (2-3 nt) enables both parallel and anti-parallel conformations to form [52]. Recently, we illustrated that loop permutation also influences G4 stability and topology [53]. Additionally, loop length has been associated with G4 molecularity, two short loops (1–2 nt) have been shown to favor intermolecular complexes in potassium [50]. Recently, Li and Mergny *et al.* illustrated that loop permutation is an inescapable factor to G4 stability and topology. The difference in melting temperature caused by loop permutation can reach a maximum of 17.0 °C. Sequences with a long center loop are more prone to hybrid topologies and demonstrate the greatest stability. On the other hand, sequences with a short center loop are more likely to fold into parallel topologies with lower stability [53].

vi) Flanking nucleotides: The core G4 motif may be flanked by one or more nucleotides at both extremities. The addition of flanking nucleotides favors a parallel topology, and flanking nucleotides at the 5' end exert a greater influence on topology than those at the 3' end [54]. The presence of nucleotides at the 5' end prevents a strong *syn*-specific hydrogen bond for a 5'-terminal guanine which is frequently observed in non-parallel conformations [54].

1.2.2 i-motif

i-motif structures involve two parallel-stranded duplexes, antiparallel to each other. Each duplex is composed of two or more hemi-protonated cytosine pairs ($C\cdot CH^+$), as shown by Guéron *et al.* in 1993 (**Figure 6a**) [55, 56]. Each non-terminal base pair from one duplex is intercalated between two base pairs from the other duplex, giving i-DNA its name: i- stands for *intercalated*. The helical rise between each hemi-protonated cytosine pair is short (2.8 Å), but this value corresponds to two base pairs belonging to two different duplexes: within a given duplex, this distance is doubled, and each duplex is therefore extended and underwound as compared to B-DNA [57]. The formation of hemi-protonated cytosine base pairs is

optimal at mildly acidic pH, since under more acidic conditions, all cytosine bases would be protonated. Therefore, the most suitable pH for the i-motif is between 4 and 5, close to the pK_a of cytosine [58].

According to the number of DNA strands that constitute the i-motif structure, i-DNA can be either monomolecular, bimolecular or tetramolecular. i-DNA is less polymorphic than a G-quadruplex, as the relative orientations of the strands are fixed: two diagonally arranged strands are always parallel to each other (allowing the formation of parallel C•CH⁺ base pairs) while adjacent strands are always in an antiparallel orientation (**Figure 6b**). According to the characteristics of the i-motif structure, when the outmost C•CH⁺ base pair is at the 3'-end, the structure is known as 3'E, while in the 5'E topology, the terminal C•CH⁺ base pair is at the 5'-end [58, 59].

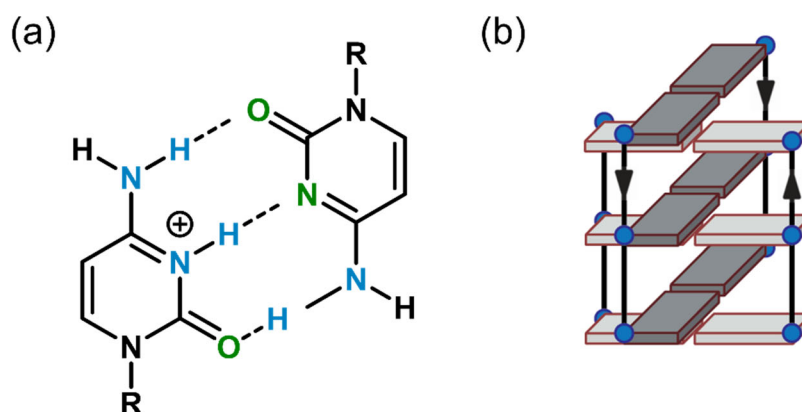


Figure 6 (a) Hemi-protonated C•CH⁺ base-pairs (donor and acceptor groups in blue and green, respectively). (b) Tetramolecular i-motifs in 5'E topologies. C•CH⁺ base pairs from each duplex are shown in different shades of gray.

As for other nucleic acid structures, temperature, cation type and concentration have an effect on the stability and conformation of i-DNA. On the other hand, while extreme pH may destabilize most conformations, i-motif stability is exquisitely sensitive to pH. Mergny *et al.* illustrated that an increase in pH of one unit results in a 23.8 °C drop in T_m [48]. As demonstrated by vibrational and electronic circular dichroism (CD) spectroscopy, Ag⁺ stabilizes cytosine base pairing and enables the formation of an i-motif-like structure at pH up to 10 [60]. Mergny *et al.* showed that increasing NaCl concentration to 100 mM destabilizes the i-motif structure, at any pH above the pK_a of cytosine. Increasing the NaCl concentration to 300 mM had no further destabilizing effect. There were no differences in i-motif stability upon addition of 5 mM Mg²⁺, Ca²⁺, Zn²⁺, Li⁺ or K⁺ in the presence of 100 mM NaCl at pH 6.4 [61]. i-DNA with relative longer central loops are more stable than those with shorter loops [62]. The nature of the bases in the loops is also important. Both T-T pairs in loops connecting C-tracts [63] and the A-T Hoogsteen base pair between two antiparallel oriented C-rich strands [64] can improve i-motif stability. Lewis *et al.* observed that a C-rich

promoter sequence fold into a stable i-motif structure in 30% w/w PEG₈₀₀₀ solution at pH values as high as 6.7 [65]. However, other crowding agents such as Ficoll70 could not stabilize i-DNA [48], which implies that PEG stabilization may not be related to molecular crowding.

1.2.3 Other four-stranded structures

Besides tetrameric complexes involving mostly guanines (G4) and cytosines (i-DNA), heterozygous G-C-G-C quadruplexes may also be formed. The bimolecular G-C-G-C quadruplex structure of d(GCATGCT) contains two G-C-G-C tetrads. It is interesting to note that d(GCATGCT)₂ form further Watson-Crick H-bonds to the complementary strand through G and C residues (**Figure 7a**), and additionally, the dimeric structure is stabilized by hexamine cobalt(III) and highly defined water molecules [66]. d(GAGCAGGT)₂ forms a head-to-head dimeric quadruplex containing sequentially stacked G-C-G-C, G-G-G-G and A-T-A-T tetrads in 1 M NaCl. Two Watson-Crick G-C pairs aligned directly opposite each other constitute the G-C-G-C tetrad (**Figure 7b**) [67]. d(GGGCT₄GGGC) is a repeated motif found in adeno-associated virus (AAV), which adopts the direct G-C-G-C tetrad alignment and forms a quadruplex [68].

Four different nucleobases can be fully integrated into one tetrad. González *et al.* first reported the self-associated intermolecular four-stranded structure consisting of two minor groove-aligned G-C-A-T tetrads, formed by a cyclic oligodeoxynucleotide dc(CGCTCATT)₂, resulting from the interaction of G-C and A-T base pairs through their minor groove sides [69].

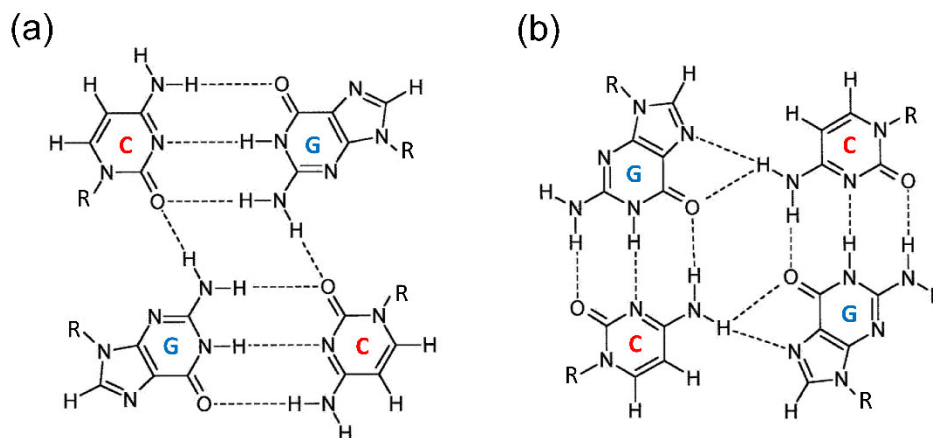


Figure 7 Schematic views of G-C-G-C quartets observed (a) between two d(GCATGCT) (PDB: 1MF5) and (b) in two d(GAGCAGGT) (PDB: 1JVC).

2. G-quadruplexes in genomes and their biological significance

Putative G-quadruplex sequences (PQS) are frequently found in telomeric regions, ribosomal DNA, the immunoglobulin heavy chain class switch recombination region, and transcriptional regulatory regions of a number of genes, especially oncogenes [70]. Telomeres (the extremities of linear chromosomes) are generally composed of short tandem G-rich DNA sequences [71, 72] and were the first motifs to be studied for G4 formation. As early as 1988, a highly conserved microsatellite repeat sequence d(TTAGGG) has been found in the human telomeric region [73] and several seminal papers demonstrated that telomeric motifs may adopt a quadruplex conformation *in vitro*. Since then, numerous other intramolecular G4 structures have been studied *in vitro*, and some of G4s can fold *in vivo* and play roles in a variety of biological processes. G4 structures may exert regulatory functions in a number of species and are also potential hurdles that need to be solved for proper replication, transcription or translation.

Two initial bioinformatics studies estimated that there were hundreds of thousands of G-rich regions with high potential to form G4 structures in the human genome [74, 75]. PQS are not randomly distributed, with numerous promoter regions in human genes associate to PQS, implicating that G4s play a role in transcription [76]. In addition, the vast majority of human replication origins were found to be in close proximity to PQS, implying that G4s also play a role in replication [77]. PQS are also present in mitochondrial DNAs [78] and chloroplast DNAs [79].

The formation of G4 is not restricted to DNA: G4s also exist in RNA molecules and are often extremely stable. Initial studies were focused on mRNA, in which G4s may affect pre-mRNA processing (splicing and polyadenylation), mRNA turnover, localization, and translation [80]. The presence of G4 motifs in non-coding RNAs (ncRNAs) suggests that RNA PQS are capable of modulating post-transcriptional gene expression *in vivo* and regulating miRNA synthesis [81]. Other ncRNAs have been reported to adopt a G4 fold: this is the case for example for two telomerase-related RNA: *hTR* and *TERRA* [82]. More recently, strong G4 motifs were found in ribosomal RNA (rRNA) [83]. Thus, G4 structures are promiscuous in a variety of nucleic acids and participate to a number of biological processes.

Some authors have proposed to correlate G4 propensity to large-scale evolutionary processes. One study predicted PQS on the genomes of 37 species at 14 representative evolutionary nodes within the eukaryotic clade of the tree of life, spanning a widely broad evolutionary scale from single-celled fungi to higher mammals and human. G4s clustered in chromosomes and were more abundant in gene bodies.

G4s with short loops were maintained in the majority of species, but loop length diversity remained as well, particularly in mammals. These observations imply that organisms may have evolved G4s into novel reversible and elaborate regulatory factors through evolution. Thus, the possible evolutionary patterns and functional implications of G4 motifs were comprehensively deciphered [84]. Together, these observations demonstrate or suggest that G4 are involved in physiological processes, such as telomere maintenance, oncogenes expression, genome instability [85]. G4 may also be important in species evolution [84].

2.1 Mapping G-quadruplexes in human genomes and other species

In 2005, Todd *et al.* [74] used bioinformatics tools to analyze potential G4 DNA forming sequences in the human genome. They defined a "classical" general consensus for G4 formation and proposed a common general sequence for G4 motifs: $G_{3-5}N_{L1}G_{3-5}N_{L2}G_{3-5}N_{L3}G_{3-5}$, where N_{L1} , N_{L2} and N_{L3} correspond to the loops of G4, which length was arbitrary limited from 1 to 7 nt and can be composed of any nucleotides, including guanines. Through predictions, the study found and located 375,157 PQS, determined the characteristics of each type of G4, and inferred their potential biology function. In parallel, Huppert and Balasubramanian outlined a similar rule for predicting putative quadruplex sequences based on the primary DNA sequence, predicting that any sequence with the $d(G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+})$ consensus would fold into a quadruplex under near-physiological conditions [75]. This definition has been used to predict, analyze and locate PQS in genomes of several species, including fungi (*i.e.*, *Saccharomyces cerevisiae* [86]) and bacteria (*i.e.*, *Escherichia coli* [87]).

In the past decade, novel bioinformatics algorithms (*e.g.*, cGcC [88]) have been developed to predict G4-forming sequences. G4Hunter [89] is an algorithm predicting the G4-forming possibility of a sequence, which takes into account *G-richness* and *G-skewness* of a DNA or RNA sequence. The G4Hunter algorithm has been integrated and packaged as a web application [90]. The association between G4Hunter application and the NCBI gene database allows to infer gene serial numbers from NCBI directly. As an advanced and easy-accessible G4 mapping tool, G4Hunter has worked for the mapping and analysis of multiple species, including viruses (*i.e.*, pathogenic virus *H1N1* [91]). I personally contributed to the experimental analysis of motifs found in all available archaeal genomes [92]. Distributions of PQS were identified in all archaeal species with highly significant differences in frequency. The presence of PQS is not random with an over-represented bias in non-coding RNAs, suggesting possible roles for ncRNAs regulation. Very recently, we also analyzed PQS density in various parasitic helminths. Stable G4 formations of multiple sequences in four different parasitic helminths were validated *in vitro* [93].

An experimental confirmation is generally required to support G4 formation, at least *in vitro*, for some of the predicted PQS in genomes. Given than hundreds of thousands of candidates may be proposed, high-

throughput experimental methods are needed. G4 chromatin immunoprecipitation sequencing (ChIP-seq) has been developed to determine G4 structures genome-wide directly in chromatin [94].

2.2 G-quadruplexes in gene promoters

Many G4 structures formed in gene promoters exhibit physicochemical and structural characteristics that suggest druggability. The "classical" or "historical" hypothesis was that G4s located in promoters would work as repressors of transcription. Bioinformatics results in the human reference genome showed that PQS are over-represented in promoters. Approximately a half of PQS in the transcriptional regulatory region (downstream 2,000 bp to upstream 1,000 bp around transcription start site, TSS) of human genome are concentrated in the proximal part, where is up to 500 nt away from the TSS [76], and more than 40% annotated genes in human promoters contain at least one PQS [95]. The promoters of human oncogenes and regulatory genes contain more G4 motifs than the promoters of housekeeping and tumor suppressor genes (TSGs) [96], suggesting that G4 structures may be involved in proliferation and transformation. The human genome has more than 640,000,000 single-nucleotide variants (SNVs), which are the most prevalent type of genetic variation in individuals or populations. When an SNV occurs within a PQS motif, it has the potential to interfere with the folding (or stability) of G4 structures. Very recently, a genome-wide survey illustrated that millions of PQS may be altered by SNVs, about 70% of the SNVs-PQS associated with genes with a profound enrichment near TSSs. Particularly, SNVs-PQS were more frequently found in oncogenes and TSGs than in most genes. This suggests that SNVs-PQS may play a role in carcinogenesis, which can be triggered by aberrant expression of oncogenes and tumor suppressor genes [97]. We list some well-characterized G4-forming sequences in cancer-related gene promoters in **Table 2**.

Table 2 Examples of G4-forming sequences in cancer-related genes.

Oncogene	Acronym	Sequence (5'-3')	Topology* (PDB No.)	Ref
<i>hTERT</i>	<i>hTERT</i>	GGGGAGGGGCTGGGAGGGCCCGGAGGGGGC TGGGCCGGGGACCCGGGAGGGGTCTGGGACG GGGCGGGG	P (-)	[98]
<i>c-MYC</i>	<i>Pu27</i>	TGGGGAGGGTGGGGAGGGTGGGGAAGG	Ap (-)	[99]
	<i>Pu24</i>	TGAGGGTGGGGAGGGTGGGGAAGG	P (2A5P)	[100]
	<i>Myc-1245</i>	TTGGGGAGGGTTTTGAGGGTGGGGAAT	P (6NEB)	[101]
	<i>Myc-2345</i>	TGAGGGTGGGGAGGGTGGGGAA	P (7KBV)	[101]
<i>KRAS</i>	<i>KRAS-21R</i>	AGGGCGGTGTGGGAAGAGGGA	P (5I2V)	[102]
	<i>KRAS-32Rh</i>	AGGGCGGTGTGGGAAGAGGGAAGAGGGGGA GG	P (6T2G)	[103]
<i>HRAS</i>	<i>Hras-1</i>	TCGGGTTGCGGGCGCAGGGCACGGGCG	Ap (-)	[104]
	<i>Hras-2</i>	CGGGGCGGGGCGGGGGCGGGGGCG	P (-)	[88]

*P and Ap refers to parallel and antiparallel G4 structures, respectively. (-) means no structure available in the PDB

hTERT: Human telomerase reverse transcriptase (*hTERT*) is the catalytic subunit of telomerase, which is often reactivated in cancer cells. *hTERT* is an important oncogene with a compact stacked three-G-quadruplex within its promoter [98]. *hTERT* G4 structure in the core promoter has been proposed to be a major factor to downregulate gene expression [105].

c-MYC: *c-MYC* belongs to the *Myc* family, which is one of the earliest known oncogenes. *c-MYC* serves as a gene-specific transcription factor involves in cell cycle, apoptosis, metabolism, cell differentiation, and adhesion. The nuclease hypersensitive element (NHE) III₁ region is located upstream of the predominant promoter 1 of *c-MYC*, which controls 85–90% of *c-MYC* transcription. Pu27, a G-rich motif in NHE III₁ region, has been reported to be biologically relevant when it forms a 'chair-type' G4. Stabilizing G4-formation (via G4-ligands or mutation) down regulated *c-MYC* expression in Ramos cells [99].

KRAS: The *KRAS* gene belongs to the Ras genes family, which also includes two other genes: *HRAS* and *NRAS*. These gene-related proteins play critical roles in cell division, differentiation, and cell self-apoptosis. The nuclease hypersensitive polypurine-polypyrimidine element (NHPPE) in the proximal promoter region of *KRAS* gene is an essential element to activate transcription in both human and mouse cells. E. Xodo *et al.* suggested that the intramolecular parallel G4 formed within the NHPPE of *KRAS* may be involved in the

regulation of transcription. A G4 ligand, TMPyP4, was proposed to break the equilibrium between double-stranded and G4 formation of NHPPE by inducing the PQS fold into a G4 structure. TMPyP4 has been shown to down regulate *KRAS* expression, possibly via the stabilization of NHPPE G4, which would interfere with the binding of RNA polymerase and inhibit transcription [103].

HRAS: As previously mentioned, *HRAS* and *KRAS* belong to the same family. The *HRAS* protooncogene can be activated by point mutations: for example, *HRAS* is frequently mutated in urinary bladder tumors and its overexpression is highly correlated with tumor invasiveness. HRAS-1 and HRAS-2 are two neighboring PQS immediately upstream of *HRAS* TSS, which repress *HRAS* transcription in a coordinated and efficient way. The two neighboring PQS work like an off-switch to regulate *HRAS* transcription; a dramatic transcription arrest is observed *in vitro* when the two G4 are stabilized [106].

2.3 G-quadruplexes in gene expression: beyond downregulation

Similar to supercoiling, which has both negative and positive effects on gene transcription, G4 structures have been proposed to influence gene transcription [107]. Some authors propose that the effect of a G4 structure may depend on which DNA strand it is located. As shown above, several well-known G4s in promoters downregulate gene expression. In contrast, G4 structures located on the template strand of some genes can recruit proteins to promote transcription. Nucleolin (NCL) is a highly specific G4-binding protein (G4BP) largely localized in nucleolus. NCL binds to a number of G4 motifs, including those found in the long terminal repeat (LTR) promoter of human immunodeficiency virus-1 (HIV-1) and serves as a transcription activator [108]. NCL acts as a molecular chaperone to facilitate G4 folding [109]. Myc-associated zinc-finger protein (MAZ), which is a G4-binding transcription factor binds to the promoter of murine *KRAS* gene. The G4-forming sequence in the *KRAS* promoter overlaps the binding site of MAZ [102]. Stabilization of the G4 DNA structure favored MAZ binding within the *KRAS* promoter and activated transcription, whereas point mutations disrupting the G4 conformation downregulated *KRAS* expression [110].

A bioinformatic survey showed a strand bias in the distribution of PQS in the human genome: downstream of the TSS, PQS are concentrated in the first intron and on the non-template strand, indicating that the transcribed RNA may also contain a large number of G4 sequences, which may also have an impact on RNA fate (splicing, translation...) [111]. An *in vitro* study demonstrated how G4 structures regulate Bcl-X (the dominant regulator of programmed cell death in mammalian cells) pre-mRNA alternative splicing. A G4 ligand, ellipticine GQC-05, causes splicing to switch from the dominant anti-apoptotic Bcl-X_L isoform to the pro-apoptotic Bcl-X_S isoform [112]. GQC-05 has been reported to bind to a DNA G4 structure in the *MYC* promoter and induce apoptosis in Burkitt's lymphoma (CA46) cells [113]. Relatively high densities of PQS in the first intron and on the non-template strand are conserved from frogs to humans [111]. Interestingly,

although a strand-bias exists in the genome of several species, there is no obvious asymmetry in the PQS between the non-template and template strands in *Saccharomyces cerevisiae* (budding yeast) genome [86].

G4s also play important roles in the non-coding part of RNAs. Huppert *et al.* searched PQS in and around more than 30,000 annotated UTRs of mRNAs (*Homo sapiens*). The distribution of PQS in UTRs is asymmetric, with a higher density of PQS in 5'-UTRs than in 3'-UTRs. A positional bias in two distinct regions has also been reported. A further increase in PQS density was found at the 5'-end of 5'-UTRs, suggesting that G4 may relate to transcription initiation. In contrast, PQS tend to cluster immediately after the 3'-end of genes, particularly when another gene is nearby, suggesting that G4s may act as pause elements to promote transcriptional termination, cleavage, and polyadenylation [114]. Based on cross-linking immunoprecipitation sequencing (CLIP-seq) results, Ghanem *et al.* observed that RNA-protein binding sites (RBPs) are also enriched in PQS-rich UTR regions, and that the frequency of overlapping RBPs and PQS peaks was almost 6-fold higher in the 5'-UTR of human erythroleukemic (K562) and human hepatoma (HepG2s) cells [115]. This analysis suggests that G4 formation within UTRs may also affect functions of RBPs and contribute to post-transcriptional regulation.

2.4 G-loop: G-quadruplexes co-exist with R-loops

R-loops are non-canonical three-strand secondary structures, in which one strand in a DNA duplex is melted and annealed to an RNA, generating a hybrid DNA-RNA duplex, while the other DNA strand is displaced out. Generally, formation and stabilization of R-loops are facilitated by the thermodynamic advantage of the hybrid between the nascent RNA and the DNA template strand over that of the duplex with the respective DNA strand. Besides the sequence-dependent intrinsic superior stability of the RNA-DNA duplex over its DNA-DNA counterpart, other factors may provide an advantage: breaks in the non-template DNA strand, negative supercoiling which can facilitate DNA unwinding, non-canonical DNA structures at the non-template DNA strand (*i.e.*, G4s **Figure 8** and triplexes) or sequestration of the non-template DNA strand by a ligand [116]. R-loops relate to genomic instability and replication, and their overabundance have been evidenced in a number of neurological syndromes and cancer [117].

The co-existence of G4 structures at the non-template DNA with a stable RNA/DNA hybrid R-loop was first observed in *Escherichia coli* in 2004; this novel structure was called a G-loop [118]. The presence of persistent G-loops may affect gene regulation, mRNA translation, and increase collisions between replication and transcriptional machineries, resulting in deleterious transcription-replication conflicts (TRCs) [119]. Some G4 ligands, such as TMPyP4 [119], PDS [120] and CX-5461 [121], have been shown to induce R-loop-mediated DNA damage and cell death in cancer cells. A recent study demonstrated that G4 on the non-template strand increase mRNA production rate and yield through enhancing R-loop formation. Surprisingly, the

increased elongation is a result of transcription-induced R-loop formation, which in turn favors G4 formation on the non-template strand. Therefore, increased transcription is caused by the G4-stabilized R-loop *via* a mechanism involving successive rounds of R-loop creation [122].

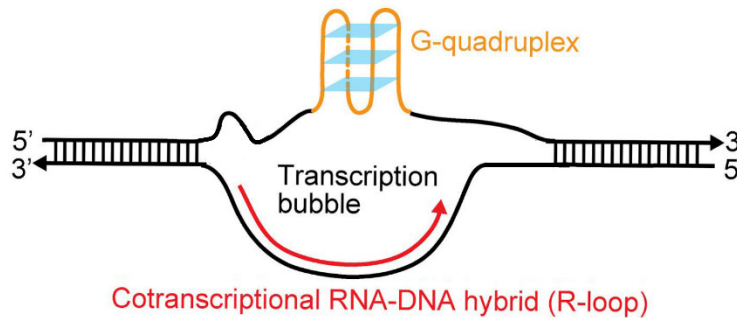


Figure 8 Scheme of the co-existence of a R-loop and a G4, giving the called G-loop.

2.5 G-quadruplexes in gene replication

G4 structures are found at the 5' ends of genes, nucleosome-free regions and inside CpG islands and in close proximity [123]. G4s have also been demonstrated to be associated with the great majority of the replication origins. Folding and unfolding of G4 structures may play an important role for gene replication, they may cause or stabilize DNA duplex unwinding in negatively supercoiled DNA, hence enabling the loading of the origin recognition complex (ORC) and other replication components [77]. High-throughput sequencing of short nascent DNA strands in human fetal lung cells (IMR-90) revealed that the majority (67%) of the 250,000 human replication origins are close to classical G4 motifs. This proportion increases to 91% when using a relaxed definition of a PQS [77].

G4 effects on DNA replication can be classified into two categories: on the initiation of DNA replication and on the process of DNA replication. An *in vivo* study concluded that G4 motif orientation determines the position of the replication start site, and decreased G4 stability impairs origin function, while the regulation mechanism still needs to be characterized [124].

During DNA replication, the two strands of the DNA double helix are unwound by helicases. However, due to the directionality of DNA polymerases, the replication process is asymmetric: the parent strand running in the 3' to 5' direction toward the duplication fork is called the *leading* strand. The replication of the leading strand is said to be continuous since DNA polymerase always runs antiparallel on the leading strand in the 5' to 3' direction to build a new complementarity strand. The other parent strand running in the 5' to 3' direction is called the *lagging* strand, its replication is discontinuous due to the mono-directionality of polymerase. A primase binds to the lagging strand at several sites and synthesizes a series of short RNA

fragments, which will serve as a template for DNA polymerase, generating the so-called Okazaki fragments [125]. This asymmetric duplication process leaves the lagging chain in a single-stranded state for a longer time than the leading strand, and the former will therefore be intrinsically more prone to G4 formation (**Figure 9**) [126]. Pif1, a 5' to 3' DNA G4 helicase, was employed to show how lagging strand G4s delayed replication in yeast. In individual yeast cells, replication rates through specific lagging strand G4 sequences *in vivo* was significantly decreased in the absence of Pif1. In contrast, replication rates through the same G4s on the leading strand were not affected under the same conditions, suggesting that Pif1 is essential only for the efficient replication of the G4s in the lagging strand.

G4s are associated to higher genomic instability when they are located on the leading strand, possibly because there are less check points due to the replication process speed [127]. G4s have been demonstrated to interrupt DNA synthesis in the absence of the DNA repair protein REV1 in a chicken B cell line (DT40), which results in localized, stochastic loss of parental chromatin marks and changes in gene expression. Histone modifications around the TSS are also strongly correlated with the position of the G4s [128]. CEB1, the G4-forming human minisatellite, was well tolerated on a leading strand but exhibits significant genetic instability when cells are treated with a G4 ligand [127].

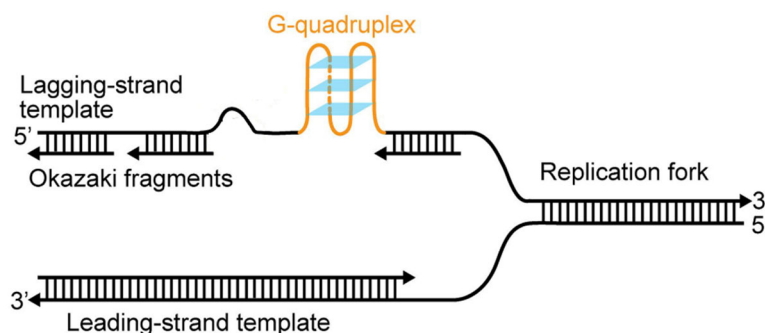


Figure 9 G4s may interfere with the replication of the lagging-strand [126].

3. Characterization of G-quadruplexes *in vitro*

A number of potential G-quadruplex motifs can be found in most, if not all, genomes. However, having a sequence that may adopt a quadruplex fold does not always mean that this quadruplex would be formed *in vitro* and in cells. In here, we briefly introduce different biophysical methods which have been included in our "G4 characterization box". Our previous experience has demonstrated that one could reach contradictory conclusions depending on the technique used [89]; thus, we recommend to use a combination of approaches to validate G4 formation.

There are a few trivial considerations to be discussed when presenting the different experimental approaches. First, we have systematically chosen to express DNA concentration as the concentration of oligonucleotide strand rather than in nucleotides, as the secondary structures are constructed by strands, not by individual nucleotides. Strand concentration may be important when considering structures of different molecularities. Its precise determination may not be as trivial as it seems, as the extinction coefficient (often provided by the manufacturer) may not be very accurate for purine-skewed sequences. In addition, one needs the sample to be unfolded for this extinction coefficient to be applicable. Denaturation may be facilitated by preincubation at high temperature (95 °C) for a few minutes (depending on the length of the strands) in the absence of salt. On the other hand, the acquisition of absorbance or CD spectra of the folded form requires that the sample is properly folded. Pre-annealing of DNA/RNA in the buffer of interest, followed by either slow or very rapid cooling, should allow the oligonucleotides to adopt thermodynamically- or kinetically-favored structures, respectively.

3.1 UV-visible absorbance spectra

UV-visible absorbance is a simple and often overlooked experimental technique to analyze nucleic acid conformations. The spectroscopic properties of DNA/RNA depend on their nucleotide content, on the properties of the solvent, and the structure adopted. Absorbance in the far UV region (around 190 nm) arises from the sugar-phosphate backbone, while longer-wavelength absorbance (200 to 300 nm) is entirely attributable to electronic transitions of purine and pyrimidine bases [129]. In the double-stranded helix, base pair stacking leads to a decrease in UV absorbance at 260 nm. This phenomenon is called hypochromicity, whereas hyperchromicity refers to the opposite phenomenon in which absorbance is increased.

The characteristic UV absorption features of quadruplexes are different from those of the double helix. Guanosine has two π - π^* transitions at 245 nm and 275 nm [130], while the formation and dissociation of intramolecular G-tetrads can be precisely monitored by absorbance changes at 295 nm [131]. Based on the characteristic transition around 295 nm, there are three biophysical ways to characterize G4 structures *in vitro* discussed below.

3.1.1 Isothermal differential spectrum (IDS)

As discussed above, different cations can stabilize G4 structures; they are actually required for proper G4 folding: in the absence of these cations (*i.e.*, in pure water or in a buffer containing non-stabilizing cations) most G-rich sequences should remain unfolded, and the addition of a favorable ion should promote folding. One may then be able to record the absorbance of the folded and unfolded species at the same temperature.

The first absorbance spectrum is therefore recorded under conditions that do not promote G4 formation (basic buffer only, no favorable ions such as Na⁺ or K⁺); alkali metal are then added, and a second spectrum is recorded after a suitable incubation time (generally short for intramolecular structures; 5 minutes should be sufficient, but much longer incubations may be required for tetramolecular complexes). The difference between these two spectra is called an *isothermal difference spectra* or IDS (proper correction for dilution must be applied for the second spectra, as the volume of the added salt is not negligible; typically, 5 to 10%).

$$\text{IDS} = \text{Spectrum in the absence of ion} - \text{Spectrum in the presence of ion}$$

The IDS of a G4-forming sample has a specific shape, with a negative peak at 295 nm. For most sequences studied in the lab, we chose to perform IDS at 25 °C / room temperature but nothing prevents the experiment to be performed at physiological temperature or, for unstable structures, at 4 °C. We use potassium to induce G4-formation and set the final potassium concentration to 100 mM; the majority of G4 structures should be stabilized by 100 mM K⁺.

3.1.2 Thermal differential spectrum (TDS)

TDS corresponds to the difference between the absorbance spectra at high (95 °C) and low temperature (typically 25 °C) in a buffer favorable for G4 formation. Usually, the samples are kept in a lithium cacodylate buffer supplemented with 100 mM K⁺.

$$\text{TDS} = \text{Spectrum at 95 °C} - \text{Spectrum at 25 °C}$$

Since high temperature tend to unfold secondary structures, the spectrum taken at 95 °C should correspond to the unfolded species, while the lower temperature spectrum (typically recorded at 25°C) may correspond to the folded species. A negative peak at 295 nm suggests G4 unfolding at high temperature.

3.1.3 General considerations on TDS and IDS

Both IDS and TDS are based on the same principle:

$$\text{TDS / IDS} = \text{Spectrum } \textit{unfolded} - \text{Spectrum } \textit{folded}$$

TDS (and to a lesser extent IDS) are now widely used methods to characterize G4 structures *in vitro*. IDS and TDS share some properties, but are not identical: while both TDS and IDS of a G4-forming sequence exhibit a negative peak around 295 nm, but significant differences will be observed at shorter wavelengths. This results from an often-overlooked consideration: the unfolded species in both differential spectra are actually different, as they correspond to high temperature or low salt conditions. We also know that the absorbance

properties of the folded and unfolded species depend on temperature, and this property will “pollute” the TDS. The following comments regarding these two differential spectra should be considered:

i) The absence of a negative peak around 295 nm is often a strong argument against G4 formation [132] (exceptions will be commented in the next §). However, the opposite is not true: a hypochromism at 295 nm is not specific for G4 formation, as other structures exhibit the same trend. Mergny *et al.* presented the TDS of several DNA secondary structures detailed in **Table 3** [132]. Besides G4 structures, other high-ordered structures (*i.e.*, i-DNA and pyrimidine triplexes) and some non-B duplexes (such as Z-DNA and parallel Hoogsteen duplexes), also lead to a negative peak at 295 nm. Thus, G4 formation cannot be validated based on this feature only. However, the overall shape – not only the region around 295 nm – of TDS seems to be specific for G4, allowing to discriminate between all these structures based on IDS or TDS only.

Table 3 Significant TDS peaks of secondary structures [132]

Secondary structures	Hydrogen bonds	Strands orientation	Significant TDS peaks*
B-DNA; AT-rich	Watson-Crick	Antiparallel	+260 nm, -280 nm.
B-DNA; GC-rich	Watson-Crick	Antiparallel	+238 nm, +278nm.
Z-DNA	Watson-Crick	Antiparallel	+ 240 nm, + 275 nm, -295 nm.
Parallel AT duplex	Reverse	Parallel	+ 260 nm, - 280 nm.
	Watson-Crick		
Pyrimidine triplexes	Watson-Crick	Hybrid (2+1)	Variable positive peak in 240 to 275 nm, - 295 nm.
G4	Hoogsteen	Variable	+ 240 nm, + 275 nm, - 295 nm.
i-DNA	C•CH ⁺	Two duplexes intercalated in antiparallel orientation	+ 240 nm, - 295 nm.

* '+' and '-' corresponding to local maxima and minima of the TDS, respectively.

ii) While the absence of a negative TDS or IDS signal at 295 nm generally indicates that the sample is not folding into a G4 structure, there are several exceptions to this rule. As both TDS and IDS correspond to the difference between unfolded and folded states, one has to check that the denaturing conditions (no salt or high temperature) are sufficient to unfold the G4 structure. Indeed, some very stable G4s may fold even in the presence of trace amounts of metal ions, and/or resist boiling (e.g., the T_m of 19wt in 100 mM KCl is higher than 90 °C [133]). As a result, the “denatured” state remains folded, and the TDS/IDS are affected. This artefact may be identified by recording the CD spectra of the “unfolded” state: if the signal is intense under these conditions, the structure cannot be considered unfolded.

iii) One should select a buffer that does not absorb light in the far-UV region (Tris is often unsatisfactory

at wavelengths below 240 nm). Cacodylate (pKa = 6.14) and acetate (pKa = 4.62) are appropriate choices when working at neutral or slightly acidic pH. They offer the additional benefit of having a nearly temperature-independent pKa [132], meaning that the pH of the solution will not change much with temperature: while a moderate pH change should not be a major issue for G4 structures, it will strongly affect the stability of other conformations such as pyrimidine triplexes or i-motif. We generally use 10 mM lithium cacodylate (Li CaCo) for near-neutral conditions, prepared from cacodylic acid and lithium hydroxide. In addition, chloride salts (UV cutoff: 205 nm) should be avoided or replaced with perchlorate (ClO_4^-) or fluoride (F^- ; not compatible with Li^+) counterparts if information in the short-wavelength region (190–210 nm) is critical.

3.1.4 UV-melting

A UV-melting experiment is commonly used to determine the stability of a complex. When a sample is heated, its absorbance properties change, indicating a conformational change in the molecule(s) in solution [134]. Heating disrupts hydrogen bonds in nucleic acid secondary structure and affects absorbance properties, which allows us to follow secondary structures' stability. The melting temperature (T_m) of secondary structures refers to the temperature at which the oligonucleotide is 50% unfolded. Obviously, a high T_m value corresponds to a high thermal stability and *vice-versa*. The absorbance at 295 nm allows us to follow folding and unfolding process of quadruplexes. A review published in 2003 detailed experimental procedures and data analysis [134]. We suggest performing UV-melting with $\approx 3 \mu\text{M}$ pre-folded samples (longer oligonucleotides will be studied at a lower strand concentration, in order to keep nucleotide concentration and absorbance in an acceptable range). As the precise absorbance measurement range of normal UV-vis spectroscopy is between 0.1 - 2.0, too low or too high concentration may lead the absorption out of this range. A typical temperature range can be 4 °C to 95 °C, but may be shrunk to a smaller interval if one has indications on the stability of the structure: there is no need to go below room temperature if the structure is expected to be very stable. We summarize the additional items that should be taken into consideration in the following:

- i)** In itself, an inverted transition at 295 nm is not indicative of G4 dissociation: other structures (pyrimidine triplexes, i-DNA) will also lead to a decrease in absorbance upon melting at this wavelength. T_m determination will then give an indication on the thermal stability of the structure but not on its nature.
- ii)** What would constitute a strong indication that this structure is actually a G4 is if its T_m depends on the nature of the cation. Performing melting experiments in KCl, NaCl, and LiCl may give precious clues: for a quadruplex, $T_{mK} \geq T_{mNa} > T_{mLi}$. Quantitative differences depend on the nature of the quadruplex: for example, for some RNA, the T_m in Na^+ may be very close to the T_m in Li^+ , but stability should be much higher in K^+ .
- iii)** For some extremely stable G4s (or if the G4 structure has been stabilized by ligands), performing UV-

melting in 100 mM K⁺ may be inadequate to observe melting or determine T_m . Lowering potassium concentration to 1, 5 or 10 mM may help.

iv) The solutions should be degassed prior melting to avoid the formation of air bubbles upon heating. If the sample is not heat-sensitive, this can be easily achieved by briefly heating the solution, then gently remove the bubbles formed (longer heating would lead to evaporation). Otherwise, degassing in a vacuum chamber or sonication (during at least 5 min) are convenient alternatives.

v) A thin layer of mineral oil on the solution in cuvettes has been proposed to solve evaporation issues. We seldom use this trick anymore, as a tight cap on the cuvette may be sufficient to prevent evaporation.

vi) Condensation at temperatures below room temperature (≤ 20 °C) can be minimized by blowing a stream of dry air on the outer cuvette surfaces [135].

vii) T_m should depend on concentration for intermolecular complexes, but remain concentration-independent for intramolecular structures. The analysis of T_m as a function of strand concentration may therefore provide additional information on the nature of the complex, as long as a wide-enough range of concentrations is used (a 5 to 10-fold change is recommended) and that the transition is clear enough to allow a precise T_m determination.

3.2 Circular Dichroism (CD) spectrum

Circular dichroism is largely empirically used in the study of the secondary structures of biomacromolecules, mostly proteins and nucleic acids. CD is relatively sensitive, allowing to study nucleic acids at strand concentration in the micromolar range, *i.e.*, lower than the one required for NMR studies. The CD spectra are particularly used in monitoring changes in a secondary structure during titration, binding, or thermal denaturation experiments [136]. In brief, any change in chirality can be followed by circular dichroism.

CD is especially useful to study G-quadruplexes, as this technique not only provides information on the nature of the structure (G4 or not), but also on the topology of the quadruplex, provided sufficient precautions are taken to exclude alternative structures that would show peaks in the same wavelength range (see below). In addition, provided a rigorous quantitative analysis is performed, it may also provide information on the number of quartets formed (see below). This sensitivity stems from the sequential arrangement of quartets, which depends on the topology. To understand stacking between two quartets, one can think of two coins stacked together (**Figure 10**): they can be head-to-tail (H-to-T), head-to-head (H-to-H), or tail-to-tail (T-to-T) (**Figure 10b**). Every stacking interaction, when more than two quartets are involved, may also be H-to-T, or H-to-H, or T-to-T. The following convention may be used: the “head” side G-tetrad corresponds to the side in which the hydrogen bonds run clockwise from donor to acceptor, while seen from the “tail” side they are counter-clockwise [137]. Generally, all guanosines in parallel G4s adopt an

anti conformation and homopolar H-to-T stacking, with a positive ellipticity at 260 nm and a negative one at 240 nm (**Figure 10c**). In hybrid G4s involving both H-to-T and H-to-H arrangements, the characterized CD pattern also exhibits negative and positive bands at around 245 and 260 nm, respectively, with an additional predominant positive band at around 295 nm (**Figure 10d**). This extra positive band probably arises from the stacking of G-tetrads with alternating polarities, as the combination of *anti* and *syn* glycosidic bond angles (GBA) alternate the polarity of G-tetrads [138]. G-tetrads adopt H-to-H and T-to-T stacking in anti-parallel quadruplexes, leading to a CD pattern far different from the one of parallel G4s, with a negative band at 260 nm and two positive bands at 240 and 295 nm (**Figure 10e**). Note that the CD signal above 220 nm exclusively results from the bases, not from the backbone. For example, the self-arrangement of guanosine nucleotides into a "G4-like" structure generates a strong CD signal, resulting from the stacking of multiple quartets which are not covalently linked through a sugar-phosphate backbone [139]. It should be noted that the CD spectra are less solid to evidence G4 forming (especially for parallel G4s), as several other structures also show CD signal in 260-270 nm.

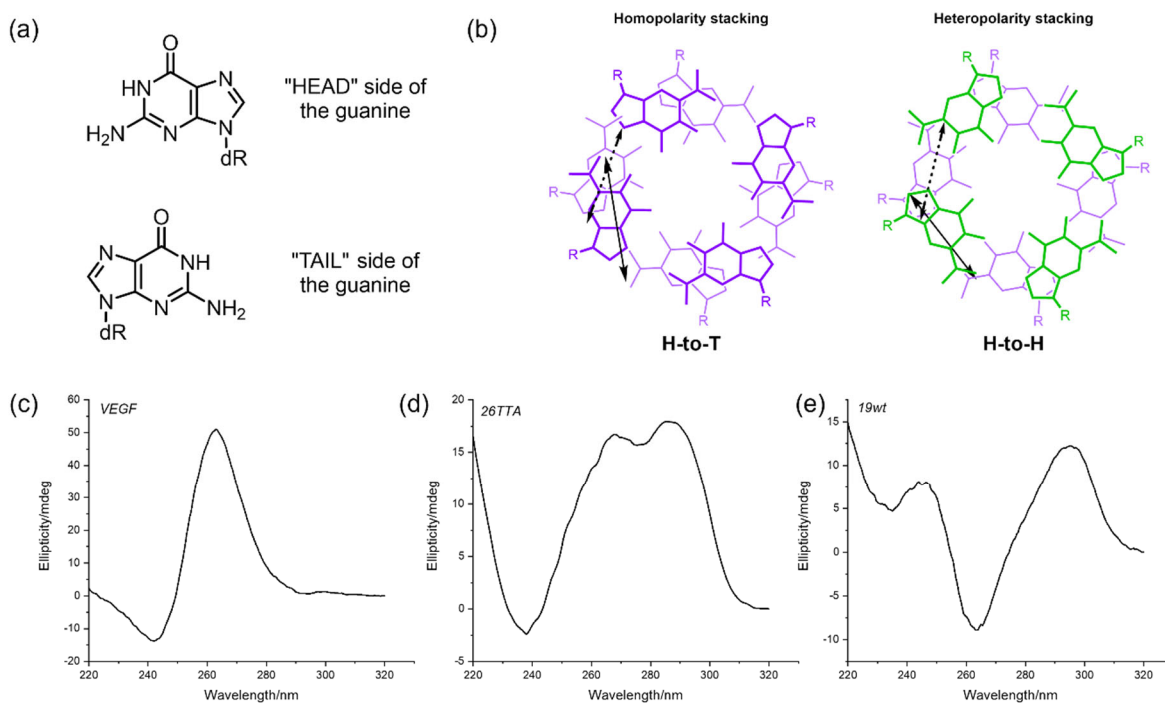


Figure 10 (a) Schematic depictions of the 'head' and 'tail' sides of the guanine. (b) Top view of the homopolar (head-to-tail) and heteropolar (head-to-head) and stacking of two G-quartets. The "head" and the "tail" sides of G-tetrads are represented in violet and green, respectively. The double-head arrows represent the transition moments corresponding to the absorption at 250 nm. CD spectra of 5 μ M pre-folded (c) VEGF (d)[CG₄CG₃CCTTG₃CG₄T], PDB: 2M27, parallel G4) and (d) wtTel26 (d[(TTAGGG)₄TT], PDB: 2JPZ, hybrid G4) in 100 mM KCl, 10 mM LiCaCo buffer, pH = 7.2; (e) 19wt (d[G₅(AG₄TAC)₂AG₄], PDB: 6FTU, anti-parallel G4) in 100 mM NaCl, 10 mM LiCaCo buffer, pH = 7.2. CD spectra were recorded on a J-1500 spectropolarimeter

(JASCO) at 25 °C, using a scan range of 320–220 nm, a scan rate of 100 nm/min and averaging four accumulations.

Raw CD spectra are expressed in ellipticity θ [mdeg] versus wavelength. A useful operation is to properly normalize CD for pathlength and concentration, which can be expressed as molar circular dichroism $\Delta\epsilon$ [$\text{cm}^{-1} \text{M}^{-1}$] or as molar ellipticity. One can use the formula: $\Delta\epsilon = \theta / (32980 \times c \times l)$ to perform this conversion, where c is the oligonucleotide concentration (in M) and l is the optical length (in cm). Proper normalized amplitudes are essential for quantifying G-quartets stacking. For parallel G4s, the CD peak around 265 nm amplitude correlates with the number of G-tetrads in intramolecular quadruplexes [140]. In addition, as the monomer, dimer and interlocked G4 structures have different numbers of stacking interactions (monomer < dimer < interlocked), the CD amplitudes at 265 nm may provide clues on molecularity, as higher amplitudes may be found for higher-ordered structures [141].

A variety of structural factors (loop types [142], sugar-phosphate backbone, strand orientation, and chemical modifications of guanines) may affect CD spectra, because they have an impact on quartet stacking geometry: CD of G4 structures is primarily determined by the stacking arrangements of guanine base steps within the G-tetrad stacks. Rather counter-intuitively, an all-parallel G4 not always gives a “parallel-type” CD spectra, with a strong positive peak at 260 nm, as the terminal quartet may flip to an all-*syn* conformation. For example, the four 5'-terminal Gs of $d[\text{TG}^{\text{Me}}\text{GGT}]_4$ (MeG = 8-methylguanine) adopt an all-*syn* G-tetrad, altering quartet arrangement and change in CD spectra without changing strand polarity [143]. This again illustrates the fact that CD primarily reflects quartet arrangement, and only indirectly strand orientation.

Specific features of CD spectra of various secondary structures are listed in **Table 4** [144]. It is important to emphasize that at first sight, the specific CD spectrum of a parallel G4 is relatively similar to the spectra of an A-form duplex (dsRNA or DNA-RNA duplex). Due to the differences in sugar pucker of a ribonucleotide, most guanosines tend to adopt an *anti* conformation and, as a consequence, the vast majority of RNA G4s are all parallel, with a CD spectrum resembling A-form RNA duplex. Finding a positive peak at 260 nm for an RNA sample is therefore not sufficient to conclude that it is forming a quadruplex, and inaccurate conclusions may possibly be reached based on CD only. On the other hand, differences in ellipticities between potassium or lithium-only conditions may be interpreted as an evidence for G4 formation. Remember, however, that traces of potassium (< 1 mM) may be sufficient to allow RNA G4 formation, and that it may be experimentally difficult to ensure that the buffer is entirely devoid of any “G4-friendly” cation. In any case, significant differences between RNA duplexes and parallel G-quadruplexes may be found at shorter wavelengths (around 200–220 nm). Nevertheless, an alternative method is strongly recommended to confirm G4 formation with an RNA sample [144].

Table 4 Significant CD peaks of secondary structures [144]

Type of conformations	Rotation / stack of stands orientation	Nucleobase orientation	Significant peaks*
B-form	right-handed	<i>anti-anti</i>	- 245 nm, + 260 to 280 nm.
A-form	right-handed	<i>anti-anti</i>	-210nm, +260nm.
Z-form	left-handed	<i>anti-anti</i>	+ 260nm, - 290 nm.
Parallel G4	parallel	<i>anti-anti-anti-anti</i>	- 240 nm, + 260 nm.
Antiparallel G4	antiparallel	<i>anti-anti-syn-syn/ anti-syn-anti-syn</i>	+ 240 nm, - 260 nm, + 290 nm.
Hybrid G4	hybrid (3+1)	<i>anti-anti-anti-syn / syn-syn-syn-anti</i>	- 245 nm, + 260 nm, + 295 nm.
i-DNA	antiparallel**	All <i>anti</i>	+ 290 nm.

* "+" and "-" corresponding to positive and negative peaks, respectively.

** referring to consecutive base pairs in the structure, which belong to two different duplexes.

3.3 Nuclear magnetic resonance (NMR)

NMR is a physical phenomenon in which the nuclei of particular isotopes behave as minuscule magnets and process around a static magnetic field that is applied externally. The frequency of the precession, also known as the "resonating frequency," is determined by the chemical environment around the nuclei, the isotope, and the strength of the external magnetic field. The resonant frequency of a nucleus of a given isotope in a given static magnetic field is only determined by the chemical surroundings of the nucleus in a molecule. As a result, measuring the resonating frequency of the nuclei in a molecule allows us to characterize the molecular chemical environment. NMR has been widely applied in structural studies of tiny organic molecules and biomacromolecules in solution.

While a complete structural determination is beyond the scope of this §, recording a simple ^1H NMR *in vitro* may suffice to confirm or exclude G4 formation. When imino protons from guanines or thymines (or uracil for RNA) are involved in hydrogen-bonds or are protected from exchanging with water molecules they become detectable by ^1H NMR. Depending on the type of nucleic acids base pairing these protons will resonate at specific frequencies. For instance, Watson-Crick bps CG or AT (as well as AU) usually present one peak around 12–13 ppm [145] and 13–14 ppm [146] respectively. On the other hand, Hoogsteen bps such as GG or GU (in RNA) will present two peaks in the 10–12 ppm region [147].

In the case of G-quadruplex formation, the four imino protons from guanines involved in a G-tetrad will present four peaks in the 10–12 ppm region, significantly different from the imino protons in Watson–Crick bps (12 - 14.5 ppm) [145] or in i-DNA (15 - 16 ppm) [146]. As a result, 1D ^1H NMR will easily discriminate between G-tetrads formation and CG or AT bps, but an ambiguity will remain between a G-tetrad and GG or GU base pairs. In an ideal situation, counting the number of imino peaks may allow the determination of the number of quartets involved in the structure. This is unfortunately rarely the case, as poor spectral resolution and structural polymorphism often lead to complex spectra. Additionally, NMR may also be a powerful tool to study G4-ligand interactions, cation coordination sites in G4 structures, detection of intermolecular hydrogen bonds, *etc.* A number of high-resolution NMR structures of inter- and intramolecular quadruplexes are available in the PDB (<https://www.rcsb.org/>), and provide invaluable information on G4 structure and dynamics.

3.4 Gel electrophoresis

A specific migration pattern of a DNA/RNA sample (which must be of high purity – this can be checked first by a denaturing gel) on a native (*i.e.*, non denaturing) gel may provide indication of a secondary structure formation, but rarely gives clues on the nature of the folded form. Gel electrophoresis is a frequently used technique for the identification, quantification, and purification of nucleic acids. The sugar-phosphate backbone of nucleic acids is negatively charged in neutral and alkali environments, allowing the nucleic acid molecules to migrate in an electric field from the negative (cathode) to the positive (anode) pole.

The gel mobility of nucleic acids is determined by their mass, size, shape and charges. Nucleic acids having the same number of nucleotides display a mobility which depends on a number of different parameters, and it is primarily determined by their apparent size and shape (in another words, their structure): large molecules move slowly through the gel while small molecules move faster. In principle, gel electrophoresis should provide information on molecularity. The mobility is expected to decrease as the number of strands involved increases. Intramolecular complexes should migrate faster than bimolecular or tetramolecular ones [148]. However, the situation can be complex, as apparent charges also play a critical role: nucleic acids are polyelectrolytes, around which cations may condense and effectively screen a part of the net charge affecting mobility. Relative mobilities are also affected by acrylamide content and acrylamide/bis-acrylamide ratio. Overall, the mobilities of G4s may differ from those of the corresponding single-strands, and also depend on topology. An accelerated migration may suggest that a particular structure is formed; however, it gives very little information on the nature of the folded form, and any conclusion on molecularity should be taken with precaution.

On the other hand, after performing electrophoresis, a direct way to evidence if some bands correspond to quadruplex species is to stain the gel with a G4-specific light-up probe. One can for example use specific fluorescent G4 ligands such as NMM [149], ThT [150] or cNDI-2 [151] (see § 4.1 *Small molecular G-quadruplexes ligands* for detail information). These dyes provide a direct method to identify G4 structures by gel, as these fluorescent dyes bind and illuminate exclusively G4 bands. There are two reasons why samples and dyes should not be pre-mixed before gel electrophoresis: **i**) G4 ligands may actually induce G4 formation in the test tube (*i.e.*, acting as chaperones [64]), leading to altered migration and possible false positives; **ii**) these ligands may affect the mobility of the sample in the gel. The reference bands are not available if samples and ligands have been mixed before.

Therefore, we recommend to first evidence G4 structures on a gel *via* specific staining by G4 ligands, for example, NMM is excited at 393 nm and the emission is recorded at 610 nm, or ThT with the excitation wavelength (λ_{ex}) at 425 nm and the emission wavelength (λ_{em}) set at 490 nm, then use a “general” staining procedure to evidence all kinds of nucleic acids after washing the gel to remove G4 ligands. Sybr Gold, which is the most widely-used commercial nucleic acids dye with λ_{ex} at 495 nm and λ_{em} at 537 nm, can stain all types of structures. Compared to G4-specific dyes, this broad indicator always has a high affinity for all nucleic acids and cannot be easily removed or replaced by other ligands, explaining why we recommend to use it only *after* G4 staining.

3.5 Intrinsic fluorescence

Similarly to isolated nucleosides, *ss* and *ds* nucleic acids emit fluorescence upon excitation with UV light; however, due to the emission in near-UV spectral range (300–400 nm) and low quantum yield (10^{-5} to 10^{-4}), this phenomenon is often considered insignificant. Remarkably, certain secondary structures, in particular G4s, are characterized by a strongly enhanced fluorescence, with broad, red-shifted peaks tailing up to 450 nm and quantum yields reaching $\sim 3 \times 10^{-4}$, which is at least 3-fold higher than the corresponding single-strands [152]. The fluorescence properties of G4s arise from the formation of excimers in guanine residues defining the core of a G4 structure [153, 154]. i-Motif structures also show enhanced fluorescence, although the difference with respect to single-strands and duplexes is more subtle.

Due to low fluorescence quantum yields of nucleic acids, measurement of their intrinsic fluorescence requires a spectrofluorometer with a sufficient signal-to-noise ratio (at least 2000 : 1 in the near-UV range, typically measured using the Raman scatter peak of water). While this performance is readily available with high-quality research instruments, entry-level spectrofluorometers may be inappropriate for using this method [155]. In addition, care must be taken to avoid the traces of fluorescent impurities in water and salts

used for the preparation of buffer solutions, as well as in oligonucleotides: that should have at least reversed-phase high-performance liquid chromatography (RP-HPLC) purity grade.

3.6 Fluorescence “light-up” assays

Fluorescent ‘light-up’ probes are important tools for G4 characterization *in vitro* and possibly *in vivo*. G4-specific fluorescent dyes with excellent spectroscopic properties are absolutely critical for this assay. The distinct physical properties between single-strands, *ds* helices and G4 structures offer the possibility of structure-selective binders. Regarding *in vitro* applications, their increased fluorescence emission in the presence of a G4-forming sequence may be used to evidence G4-formation in a test tube, microwell plate, or in non-denaturing gels. A large number of fluorescent small molecular compounds that display strong fluorescence enhancement upon G4 binding have been described [156, 157]. We therefore summarize representative fluorescent G4 ligands (some of them are commercially available and can be easily accessed) from typical chemical families. Excitation and emission wavelengths and topologies preferences of G4-specific ligands are summarized in **Table 5**, and further information (chemical structures, binding mechanisms and biological applications) are discussed in § 4.1 *Small molecular G-quadruplexes ligands*.

Table 5 Summary of G4-specific fluorescently ligands.

Families	Abbreviation* (CAS No.)	Excitation nm	Emission nm	Binding preference	Ref
Porphyrins	NMM (42234-85-3)	380	610	Parallel G4s	[158]
core-extended Naphthalene	cNDI-2 (-)	650	687	Parallel G4s	[151]
Diimide derivatives	cNDI-3 (-)	605	663	Hybrid and antiparallel G4s	[159]
Triphenylmethane derivatives	MG (569-64-2)	617	650	No G4 topology preference	[160]
	CV (548-62-9)	540	650	Antiparallel G4s	[161]
Benzothiazole derivatives	ThT (2390-54-7)	420	490	Possible binding to AG-rich strands and <i>ds</i> DNA cavities	[150]
	ThT-HE (1641591-70-9)	415	485	Parallel G4s	[162]
Styryl derivatives	DASPMI (2156-29-8)	450	584	Parallel G4s	[163]
	Distyryl-1p (2481650-06-8)	508	570	Antiparallel G4s	[163]

* Commercially available and most useful compounds are shown in bold; chemical structures are shown in **Figure 11**.

It should be noted that G4 ligands may alter the secondary structure of oligonucleotide strands, as NMM [164] and ThT [165] have been implicated in DNA conformational rearrangement. In other words, these probes may lead to false positives for G-rich but non G4-forming strands. The equilibrium of G4 and ligands binding is related to the concentration of each component. Besides a possible switch in topology, high concentrations of ligands may cause aggregation or even precipitation.

3.7 Additional methods of interest

Additional techniques provide interesting information to characterize these quadruplexes *in vitro*. These methods may not always directly answer the question whether a sequence is forming a G4 or not, but rather provide information on the folded structure. Size exclusion chromatography (SEC) [166] and analytical ultracentrifugation (AUC) [167, 168] allow the analysis of the folding of an unstructured single-strand oligonucleotide into a compact structure upon G4 formation. Mass spectrometry (MS) may also be used to analyze non covalent nucleic acid complexes (for a recent review, [169, 170]). MS offers unique advantages owing to its ability to directly measure strand stoichiometry, even in a mixture. In addition, advanced MS approaches such as reactive probing, fragmentation techniques, ion mobility spectrometry or ion spectroscopy [171] may provide additional information on the complexes.

In addition, even if G4 formation is clear, obtaining reliable information on its stability, molecularity or topology can be difficult in some cases. Some of the methods described here yield information on some, but not all, of these aspects. Determining a high-resolution structure is even more challenging and time-consuming. For instance, guanines involved G-tetrads may be identified using dimethyl sulfate (DMS) footprinting. The N7 position of the guanines involved in Hoogsteen bonding is protected and therefore not methylated by DMS; a subsequent piperidine treatment results in chemical cleavage of the methylated guanine residues. This DMS footprinting experiment involves oligonucleotides labeling and methylation, chemical cleavage, denaturing gel electrophoresis and imaging [172].

4. Specific G-quadruplex ligands

In 1997, the first small molecule able to interact with G4 structures was reported. This low-molecular weight molecule stabilized a G-quadruplex and inhibited telomerase activity [173]. In the following decades, more than one thousand G4 small molecular ligands have been identified [174]. A number of studies demonstrated that G4 ligands interact with telomeric G4 structures, interfere with telomeric functions, and

exhibit significant antitumor activity *in vivo*. Telomeric and telomerase G4s have been purposed as potential therapeutic targets [175, 176], and at least one G4 ligand (CX-5461) is currently being tested in clinical trials. Ligands are also utilized extensively in G4 characterization assays: the light-up assay is based on highly specific fluorescent ligands (see previous §), while the competition binding of PhenDC3, serves as the foundation for both FRET-MC and iso-FRET, which are two new methods reported later in this thesis. Some G4BPs (*i.e.*, G4 helicases) participate in cell activities. G4BPs are also involved in PQS-targeted sequencing and the immunofluorescence visualization of G4s in cells.

4.1 Small molecular G-quadruplexes ligands

The most classical mode of interaction between a G4 ligand and a G-quadruplex involves the ligand stacking on the “top” or the “bottom” external G-quartets and stabilize G4s in their initial folding topology. A recent report shows that PhenDC3 converts a hybrid human telomeric DNA into an antiparallel chair-type structure, and it inserts into two G-tetrads of the nascent antiparallel G4 [177]. PhenDC3 can also be sandwiched between two quartets belonging to two different quadruplexes [178]. Flat-shaped aromatic molecules interact by π -stacking with the external G-quartets, which is mainly controlled by hydrophobic and *van der Waals* interactions. This typical π -stacking mode has been extensively used in G4 ligands design and modifications, and several polycyclic aromatic systems involving different families have been developed for this purpose:

i) Anthraquinone derivatives: Anthraquinone derivatives have a relatively large π -surface area, which enables efficient π - π interaction with G-tetrads. The first G-quadruplex ligand and telomerase inhibitor, a 2,6-diaminoalkylamidoanthraquinone derivative, had been reported in 1997 [173]. Anthraquinone compounds with positively charged carboxamide side chains have a greater affinity towards G4 DNA over duplex DNA. Bhattacharya *et al.* designed a series of anthraquinone-contained compounds with varying lengths of side chains. These compounds have shown selective cytotoxic for cancer cells (HeLa and HEK293T) over normal cells (NIH3T3 and HDFa) *in vitro* as determined by cell viability assays [179].

ii) Porphyrins: As the “godfather” of the porphyrin family, TMPyP4 (**Figure 11a**) was discovered in 1998 [180]. It has high affinity, but little specificity [181], questioning the interpretation of the biological effects of this molecule. A very different porphyrin representative is N-methyl mesoporphyrin IX (NMM, **Figure 11b**) [182], which shows high specificity but moderate affinity [158]. Thanks to their electron deficient system, metalloporphyrins provide strong π -interaction with terminal G-quartets. Several metalated porphyrins have been developed as G4 ligands: the metal core including transition metals, such as Fe(III) protoporphyrin (known as Hemin) [183], Ni(II) and Co(III) porphyrins [184]; and main-groups metals, such as Mg(II) porphyrin [185], generating metal complex derivatives with binding constant values laying in the range 10^6 - 10^7 M⁻¹ [186].

iii) Acridines: In 2001, Neidle *et al.* described BRACO-19 (**Figure 11c**), a tri-substituted acridine G4 ligand. BRACO-19 has a binding constant of $1.6 \times 10^7 \text{ M}^{-1}$ to telomeric DNA, which is approximately 100 times greater than its binding constant for a DNA duplex [187]. BRACO-19 was reported to be an effective telomerase inhibitor [187], it decreased *hTERT* expression drastically in the human uterus carcinoma cell line (UXF1138L), and produced 96% growth inhibition against early-stage solid tumors generated *in vivo* by UXF1138L cells [188]. The compound AS1410 was obtained by modifying the substituent at position 9 of BRACO-19, 3,4-Difluorobenzylamine being replaced by 4N,4N-dimethylbenzene. Compared to BRACO-19, AS1410 has increased hydrophobicity and longer plasma half-life [189]. A novel acridine orange (AO) derivative has been developed as G4 stabilizer which binds with high affinity to KRAS22-RT G4 [45].

iv) Quinazolones: CX-3543 (commonly referred as Quarfloxin, **Figure 11d**) is a fluoroquinolone derivative that exhibits dual mechanisms of interaction to topoisomerase II (Topo II) and the G4 structure [190]. Different from other telomeric G4-binding ligands, CX-3543 induces disruption of the NCL-DNA G-quadruplex complex in the nucleolus and re-localizes NCL into the nucleus, thereby inhibiting RNA polymerase I (Pol I) transcription and inducing apoptosis in cancer cells [191]. CX-3543 was the first-in-class anti-cancer candidate as a G4 ligand, entering Phase II clinical trial in patients with low to intermediate stage neuroendocrine carcinoma as early as 2009. CX-5461 was later developed based on CX-3543. CX-5461 (**Figure 11e**) showed specific toxicity against BRCA deficiencies in cancer cells and polyclonal patient-derived xenograft models, including tumor resistant to PARP inhibition [121]. Interestingly, Uesugi *et al.* reported the synthesis of a novel quinazolone derivative, RGB1, which exhibited a reasonable affinity for RNA ($K_d = 5.9 \mu\text{M}$) *TERRA* G4, but no affinity for DNA G4s [192]. If confirmed, it would be fascinating to understand how RGB1 binds to RNA G4s only and why this binding mechanism cannot be adopted by DNA G4s.

v) Pyrido-dicarboxamide derivatives: Pyridostatin (PDS, **Figure 11f**) binds to human telomeric G4s and uncaps the human protection of telomeres 1 (POT1) protein from the telomeric 3' G-overhang [193]. A G4-forming sequence identified in the *Atg7* gene can fold into a G4 structure, and PDS treatment downregulates gene expression [194]. Due to their excellent affinity towards G4s, pyrido dicarboxamide derivatives have been modified with fluorophores and applied to light-up G4 structures directly. PDP (a derivative of PDS) was linked with a Cy5 fluorescent molecule by condensation reaction, and the functionalized compound PDP-Cy5 showed G4 stabilization properties *in vitro* and was successfully used in the detection of exogenous G4s in living HeLa cells [195]. Guillon and Mergny *et al.* described a series of pyridine derivatives, which are telomeric G4 structure stabilizers and exhibit a high selectivity for G4 over duplexes. One of them was found to be the most active against human myeloid leukemia cells (K562) [196].

vi) Bisquinolinium derivatives: 360A (**Figure 11g**) is a pyridine-dicarboxamide derivative that has been shown to stabilize G4 *in vitro*. 360A inhibits cell proliferation and induces apoptosis in telomerase-positive glioma cell lines [197]. PhenDC3 (**Figure 11h**) was developed in 2007; this molecule is locked in a H-bonded

syn-syn conformation, which makes it perfectly suitable to bind G4 structures, and it shows excellent stabilization of the human telomeric G4 sequence [44]. Due to its high G4 affinity, PhenDC3 has been proven to regulate gene expression in HeLa S3 cell lines [198]. Furthermore, PhenDC3 exhibits a low affinity for non-G4 structures such as ssDNA, ssRNA, and dsDNA [199]. Two derivatives of PhenDC3, PhenDC3-alk and PhenDC3-az, were used to investigate cellular localization of G4 drugs. Human colorectal adenocarcinoma cells (HT-29) were treated with the two PhenDC3 derivatives and the latter were post-labeled with Cy5 by copper-catalyzed click reaction. The post-labeling *via* both copper-catalyzed azide-alkyne cycloaddition (CuAAC) and metal-free strain-promoted alkyne-azide cycloaddition (SPAAC) have been performed on the two PhenDC3 derivatives. Confocal imaging highlighted that Cu/alkyne intermediates tend to localized non-specifically to nucleoli while, in contrast, SPAAC generates results that are more consistent between fixed and live cells [200].

vii) core-substituted Naphthalene Diimide (cNDI) derivatives: Like porphyrins, naphthalene diimide derivatives (NDIs) have an electron-deficient π -system, exhibiting strong electron affinity and high charge carrier mobility. The electron density on NDI planar aromatic core generates a quadrupole moment with a partial negative charge above and below the plane and a partial positive charge around the periphery. This electron-accepting core is advantageous for stacking with an electron-rich partner, such as the G-quartet [201]. Additionally, some side chain substituents (like piperazine) adopt hydrogen bonds with the phosphate groups in G4 grooves, further stabilizing the G4-NDI complex [202]. Naphthalene diimides are colorless in the absence of substituents at the aromatic core, rendering them ineffective as pigments. Harnessing the molecule with cationic chains increases affinity towards negatively-charged nucleic acids structures, but may have a deleterious effect on conformation selectivity. In fact, the most powerful ligands produce a high and possibly saturated binding response with all nucleic acid structures, regardless of their identity and conformation. Indeed, the N',N'-dimethylamino moieties of cNDI-1 are protonated under physiological conditions and are responsible for non-specific binding to DNA structures [151]. To overcome this problem, Zuffo *et al.* removed the less specific electrostatic interaction between the phosphate backbone and the protonated amine groups, and topological selectivity emerged at the expense of affinity. cNDI-2 (**Figure 11i**) with terminal hydroxyl groups preferentially binds to parallel G4s, and its fluorescence is enhanced in the presence of parallel G-quadruplexes in the low nanomolar (8-10 nM) range [151].

Antiparallel-specific binding is possible for cNDIs. cNDI-3 (**Figure 11j**) is a self-aggregating G4 ligand, exhibiting fluorescence enhancement when bound to hybrid and antiparallel G4s. Contrary to parallel G4s, ss or dsDNA addition could not change the fluorescence intensity significantly. The DNA-induced disaggregation is the key controlling factor of the light-up effect, the difference between enhanced emission (for hybrid and anti-parallel G4) and quenching (for parallel G4) may be the result of different interaction modes [159].

viii) Triphenylmethane (TPM) derivatives: Various compounds belong to the TPM family, including methyl

violet (MV), ethyl violet (EV), methyl green (MEG), malachite green (MG, **Figure 11k**), and crystal violet (CV, **Figure 11l**). These molecules have been shown to have a distinct affinity for intramolecular G4s. Affinities of TPM dyes for the intramolecular G4 (Hum21) is inversely correlated to substituent size [203]. High affinity of MG [160] and CV [161] toward Hum21 may be related to the small size of substituents. Similar to porphyrins and cNDIs, MG and CV also interact through end-stacking with the structure. High-resolution NMR spectra showed that the MG binding pocket of the RNA aptamer is highly unstructured in the absence of the ligand and forms a defined structure only upon ligand binding. The binding pocket contains an atypical quadruplex (C-G-G-A), which provides a stacking platform for MG (PDB: 1Q8N). Upon binding, the phosphate groups fold into a structure that results in an asymmetric charge distribution at the level of the binding pocket that forces the ligand to adapt through the redistribution of the partial positive charges. As homologous of MG, CV, could hardly interact to the aptamer ($K_d > 1$ mM). This result may be explained by the impossibility of the ligand to adapt well enough its highly delocalized positive charge (distributed equally across all three benzene rings) to compensate for the asymmetry of the binding pocket [204]. Later on, the end-stacking mechanism of CV-G4 interaction was demonstrated with a stoichiometric ratio of 2:1 (two CV dyes stacked on two outside G-quartets). This stacking mode increased the rigidity of the ligand and subsequently the fluorescence intensity [205]. CV has also been shown to bind to *i*-DNA and enhance fluorescence [206].

ix) Benzothiazole derivatives: Different from the examples above, benzothiazole derivatives offer a modest flat aromatic surface for stacking. Owing to their positive charges, benzothiazole derivatives are designed to become fluorescent upon binding to negatively charged nucleic acids structures *via* restricted intramolecular rotation (RIR). Thioflavin T (ThT, **Figure 11m**) is a representative example with fair, but not exquisite, G4 vs. duplex discriminating ability. ThT has been used to track G4s in polyacrylamide gels [150] and also to visualize G4 structures in cells. The natural G4 structures located in the nucleoli of living human breast cancer (MCF-7) cells are directly light up by ThT. The competitive binding between PDS and ThT showed that ThT is specific toward G4 structures in live cells, since the fluorescent foci disappear in the presence of PDS [207]. However, ThT also has been demonstrated to bind to *ds*DNA containing "cavity structures" such as abasic sites, gap sites or mismatch sites, which can provide an appropriate cavity for ThT binding [208]. ThT shows the ability to dimerize GA-rich strands in a parallel double-stranded helical structure, accompanied by over 100-fold enhancement in its fluorescence emission [209].

ThT-HE (**Figure 11n**) is a ThT derivative in which the nitrogen in position 3 of the thiazole ring was modified by an ethylamine moiety. ThT-HE showed preferentially binding to parallel G4s in both potassium and sodium buffers [210].

x) Stryryl derivatives: Preferential fluorimetric response towards G4 structures appears to be a common feature of mono- and distyryl dyes, including long-known mono-stryryl dyes used as mitochondrial probes or protein stains. The mono-stryryl dyes, DASPMI [2-(4-(dimethylamino)stryryl)-1-methylpyridinium iodide]

(DMSPMI, **Figure 11o**) or its 4-isomer, are fluorescent light-up probes characterized by high fluorometric response, excellent selectivity with respect to double-stranded DNA or single-stranded RNA controls, high quantum yield in the presence of G4 analytes (up to 0.32), and large Stokes shift (up to 150 nm). DASPMI preferentially responds to parallel G4s, and the binding of 5 μM dye to 10 μM parallel G4 in 100 mM K^+ buffer can be observed by the naked eye [163]. Distyryl-1p (**Figure 11p**) presents complementary binding preference as compared to DASPMI: The fluorescence quantum yield increases up to 550 folds when it binds to antiparallel G4, and this signal is strong enough to be observed by the naked eye [163].

As there are significant differences between the grooves of a quadruplex (**Figure 4**) and those of a duplex (**Figure 2**), grooves-binding is also worth investigating. Unfortunately, compared to π -stacking compounds, few ligands have been reported to interact with a quadruplex *via* its grooves [211]. Distamycin A was the first reported groove-binding ligand [212]. Distamycin A binds to d[TGGGGT]₄ with a 4 : 1 (ligand : G4) ratio. This ligand has a crescent shape and two ligands laid side-by-side in a medium-sized G-quadruplex groove, where the interaction between Distamycin A and the G4 structure is established by four hydrogen bonds with guanines [212].

Steric hindrance is an interesting approach to prevent ligand stacking on the G-tetrads. For example, the latter forces the metal-complexed ligand to interact with G4 *via* groove-binding. Rajput *et al.* chelated several planar polyaromatic hydrocarbon ligands with an octahedral metal, ruthenium(II) [213]. Ru(II) was protected in an octahedral environment surrounded by the ligands, preventing its stacking on G-tetrads. Unexpectedly, [(phen)₂Ru(tpphz)Ru(phen)₂]⁴⁺ and [(bpy)₂Ru(tpphz)Ru(bpy)₂]⁴⁺ were found to bind to duplex DNA as well.

The majority of G4-interacting compounds are stabilizing G4 structures; destabilization can only be caused by compounds having a higher affinity or different stoichiometry for single-strands than quadruplexes [214]. A well-known G4 stabilizer, TMPyP4, has been reported to destabilize the bimolecular G4 structure of the fragile X syndrome expanded sequence d[(CGG)₇CGTGGACTG], while stabilizing the d[GTCAGGTGC(CGG)₇]. This paradoxical effect may result from the sequence differences between these two strands: the arrangement of folded strands, groove sizes and length, and base composition of loops in different DNA G4 structures dictate their dissimilar interaction with the cationic porphyrins [214]. A recent study demonstrated the porphyrins-DNA interactions are influenced by porphyrins' charge and side chains [215].

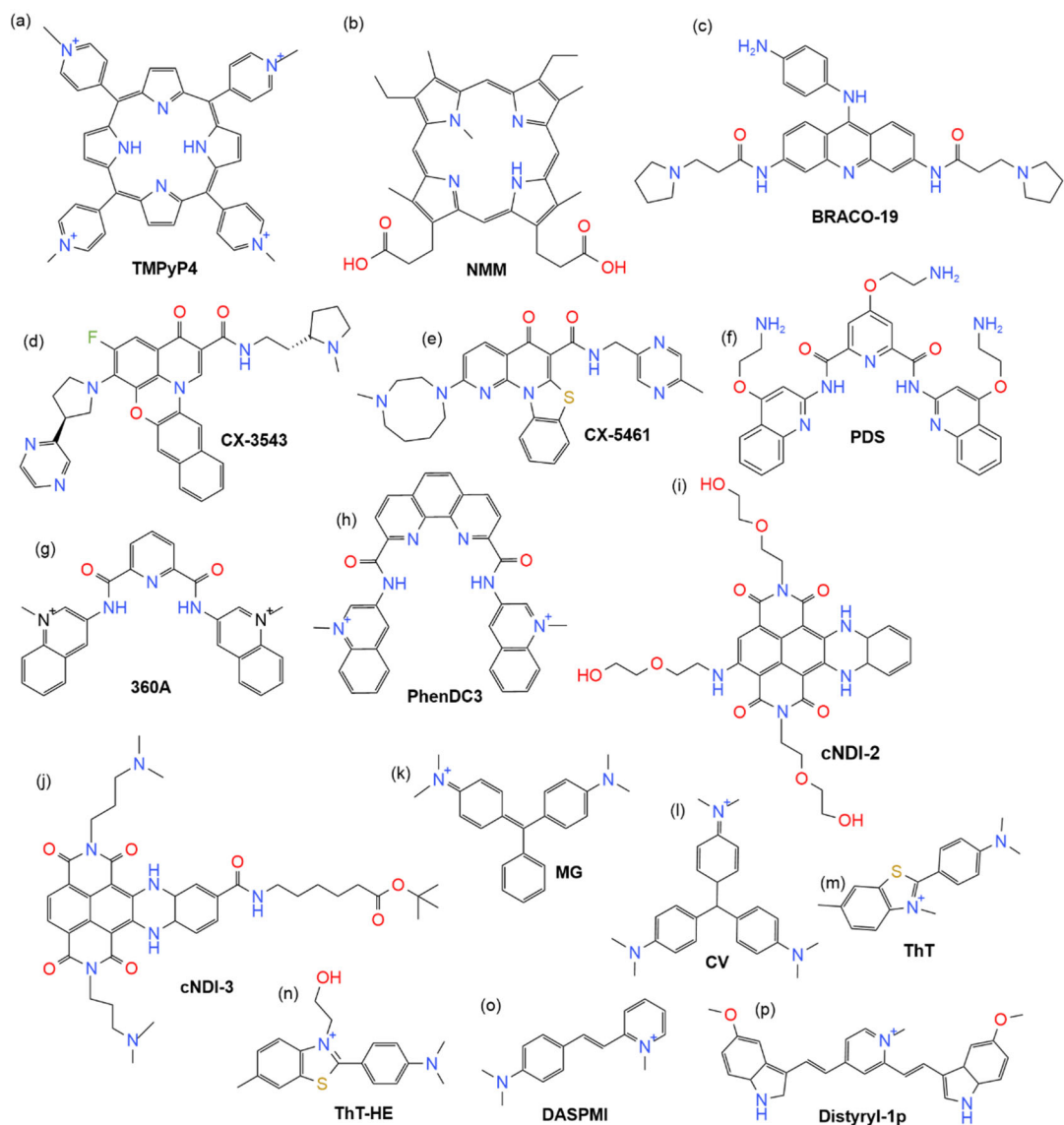


Figure 11 Chemical structures of typical G4 ligands. (a) TMPyP4 (a cationic porphyrin with modest selectivity) (b) NMM, (c) BRACO-19, (d) CX-3543 (Quarfloxin), (e) CX-5461, (f) Pyridostatin (PDS), (g) 360A, (h) PhenDC3, (i) cNDI-2, (j) cNDI-3, (k) Malachite Green (MG), (l) Crystal Violet (CV), (m) Thioflavin T (ThT), (n) ThT-HE, (o) DASPMI, (p) Distyryl-1p.

4.2 G4 interacting proteins

Several G4BPs play important biological roles. Among them, Nucleolin (NCL) is the most frequently reported G4BP for its biological functions upon G4 recognition. NCL is intimately linked to tumor proliferation and invasion; it is overexpressed on the surface of both tumor and tumor-associated endothelial cells [216, 217], promoting the development of tumor vasculature and facilitating tumor cell extravasation [218]. NCL exhibits high affinity (2.5 nM) to G4 structures [108]. As shown in § 2.3 *G-quadruplexes in gene expression: beyond downregulation*, the interaction of NCL and G4 structures in promoters influences gene transcription.

AS1411 is a man-made G4-forming DNA aptamer of NCL [219]: it binds to NCL and interferes with its cellular functions. For example, AS1411 shows antiproliferative activity against a wide range of cancer cells, and has been tested in clinical trials [220]. Besides inactivating NCL directly, AS1411 can be conjugated to act as bullet for NCL-targeted drug delivery [221].

G4 helicases include multiple eukaryotic enzymes which exploit active functions on G4s in different conditions. Helicases with G4-resolvase activity belong to extremely different families: FANCI, RTEL1, Pif1, and DHX36 (also known as RHAU or G4R1) have 3'-5' directionality, while WRN, BLM, RECQL1, and XPB have 5'-3' directionality [222]; FANCI-catalyzed G4 unwinding is adenosine triphosphate (ATP)-dependent [223], while BLM enzyme is active in the absence of ATP under physiological salt concentrations [224]. The majority of G4 helicases requires a single-stranded tail at either the 3' or 5' end to allow them loading on to the DNA substrate and act on the G4 structure [225]. One exception is the DDX5 helicase, which activates G4-unfolding without requiring a single-stranded tail. It has been shown that DDX5 directly interacts with the G4-forming *MYC* promoter region and further transactivates *MYC* transcription [226]. G4 helicases play crucial biological roles. For instance, DHX36 regulates telomerase functions by unwinding G4s within the telomerase RNA component (TERC) [227]. In addition to Pif1 (which was shown in § 2.5 *G-quadruplexes in gene replication*), FANCI has also been revealed to unwind G4 structures and support DNA replication [228]. Some studies evidenced that G4 ligands influence the activities of G4 helicases. For example, PhenDC3 limits the G4-unwinding activity of RHAU, while NMM is unable to inhibit RHAU processing [229]. The interaction between G4 ligands and helicases is still not fully understood.

Beyond the original biological applications, some G4-binding selective proteins have been isolated and applied to recognize G4 motifs. The first G4 antibody, *me^vIIB4*, was discovered in 1998 [230]. *me^vIIB4* showed affinity to promoter G4s ($K_a \approx 10^5 \text{ M}^{-1}$) and had at least 10-fold higher affinity for quadruplexes than for triplex and duplex DNA [230]. Hf2 was produced in 2008 and showed higher selectivity towards G-quadruplexes than *me^vIIB4* [231]. Subsequently, hf2 was used to pull-down G4-forming sequences from the genome of human breast cancer cells (MCF-7), providing a sequencing-based evidence to determine G4 structures existence in the B-form duplex genomic DNA [232]. 1H6 is an artificial G4-interacting protein claimed to have a high affinity and specificity for G4s. ChIP-seq technique has been used to capture G4s in living human, mouse, and chicken cells. Over 123,000 PQS have been identified in human cells [233]. Unfortunately, 1H6 also specifically binds to poly (T) DNA; therefore, the cross-reactivity of 1H6 with single-stranded T-rich DNA should be considered when interpreting 1H6 binding to (sub-) cellular structures [234]. BG4 is the most popular G4 antibody used to immunodetect G4 DNA and RNA in human cells. It has been used for immunofluorescence experiments on metaphase chromosomes [235] and the visualization of RNA G4s within the cytoplasm [236] of HeLa cells. BG4 has also been used to stain DNA G4 structures in human stomach and liver tumor tissues: the number of G4-positive nuclei in cancer patient-derived materials

increased significantly compared to the background, which implied increased G4-formation during cancer progression [237].

Globally speaking, G4 ligands and G4BPs have shown potential for G4 characterization *in vitro*, G4 gene regulation, G4 targeting and mapping in living cells, G4 staining and imaging in cells and tissues. The panel of G4-interacting molecules represents a powerful tool to study and operate G4s both *in vitro* and *in vivo*.

5. Thesis objectives

G-quadruplexes are widely distributed in genomes and play important biological roles. Bioinformatics predictions and high-throughput sequencing approaches have identified hundreds of thousands of potential quadruplex motifs, which should be experimentally confirmed. As shown in § 3.

Characterization of G-quadruplexes in vitro, several reliable biophysical methods are already available to characterize G4 structures. Unfortunately, most of these techniques are both time- and sample-consuming. For example, although some commercially available UV spectrophotometers can be equipped with multiple cells holder, allowing to process up to 8 samples simultaneously, UV-melting needs more than 14 h to acquire denaturation profiles; and a 600 MHz NMR spectrometer requires at least 350 nmol (500 μ M \times 700 μ L) oligonucleotide. To improve the signal quality, the recording time for each NMR spectra often exceed 2 h. Obviously these two techniques are not suitable for validating a large number of PQS. Differently, fluorescence “light-up” assays (§ 3.5 *Fluorescence “light-up” assays*) can contribute to PQS validation: the amount of oligonucleotide can be as low as 75 pmol (3 μ M \times 25 μ L) and the experiment can be carried out in 384-micro wells plates. However, there is no perfect assay for G4 characterization as there is no “perfect” G4 ligand with ideal fluorescent properties. Limited by the affinity of fluorescent G4-binders and by low specificity, some weakly bound G4s may be left out while some other specific structures (*i.e.*, ThT staining AG-rich duplex [209]) may be stained, generating both false negatives and false positives.

Folding/unfolding of intramolecular G4 structures can be characterized by using double-fluorescent labeled strands: the process changes the intensity of the fluorescence emission of the two molecules *via* Förster Resonance Energy Transfer (FRET), which has been widely used in characterizing nucleic acid structures (*i.e.*, hybridization [238]). FRET occurs between donor and acceptor chromophores, the energy of a donor at the electronic excited state is transferred to an acceptor through nonradiative dipole-dipole coupling, resulting in fluorescence quenching of the donor and fluorescence enhancement of the acceptor, unless the acceptor is ‘dark’ (*i.e.*, non-fluorescent) [239]. The FRET efficiency is strongly inversely correlated to the distance between the two chromophores [240, 241], which is usually limited to donor/acceptor distances of about 10 nm [242]. In an intramolecular G4, the distance between the two ends is short in the folded form, while it is longer during G4 unfolding. Therefore, the structure of a dual-labeled G4 strand (donor-G4 strand-

acceptor) can be followed by FRET [243]. As shown in **Figure 12**, the low fluorescence of the donor (FAM) at 0 °C relates to G4 formation, as the two ends of F21T are in close proximity and the FAM fluorescence is quenched *via* a FRET mechanism. With the temperature increasing, G4 unfolds and two ends are split apart, restoring gradually FAM fluorescence. On the contrary, fluorescence intensity of the energy acceptor (Tamra) decreases during the heating process. The FRET-melting curve also provides a melting temperature (T_m) value, which is an important parameter to describe G4 thermal stability. T_m is approximated by the temperature value associated to the 50% increase of the fluorescence intensity of the donor (**Figure 12**). Additionally, FRET-melting allows us to determine if a compound binds and stabilizes a G4 [244]. G4 ligand-G4 interaction often (but not always) stabilizes the G4 structure, leading to a higher T_m value.

FRET-melting can be performed with 96-microwells plates, allowing high-throughput characterization. Differently from temperature-controlled quartz cuvette holders, temperature of samples in plastic microwell plates can be changed quickly, reducing the heating time to less than 2 h. However, oligonucleotide labeling requires longer synthesis time and higher purification grade, and labeled strands are much more expensive than non-labeled counterparts. Additionally, FRET-melting cannot be easily applied to validate tetra-molecular G4s (*i.e.*, parallel tetra-molecular G4s does not put in close proximity the 5' and 3' termini).

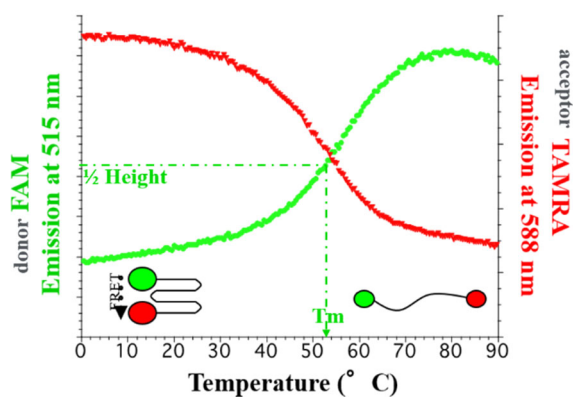


Figure 12 FRET melting curves of 200 nM F21T (FAM-d[(G₃TTA)₃G₃]-Tamra) in 100 mM NaCl, 10 mM cacodylate buffer (pH 7.0) containing 0.1M NaCl. Excitation was set at 480 nm, and emission was recorded at 515 nm (λ_{em} of FAM) and 588 nm (λ_{em} of Tamra). The figure was adopted from [243].

The main focus of this thesis is therefore to improve the throughput and the accuracy of G4 characterization methods *in vitro*, in order to provide accurate and inexpensive solutions to confirm G4 formation, even for a large number of samples. We first added unlabeled unknown competitors to a classical G4-ligand FRET-melting system. The dual-labeled stable human telomeric G4 (F21T) and an excellent G4 ligand (PhenDC3, **Figure 11h**) were employed to devise the FRET-melting competition assay (FRET-MC), and to this system a high concentration of an unknown non-labeled competitor is added to interfere with the binding of F21T and PhenDC3. If after the addition of the competitor PhenDC3 is still available to bind

F21T, higher T_m values will be measured, implying that the competitor has weak affinity toward PhenDC3, and therefore, the competitor is not a G4 structure. Conversely, if the T_m of F21T does not change in the PhenDC3-contained solution, it means that PhenDC3 forms a complex with the G4 competitor. The FRET-melting competition assay maintains the advantages of the traditional FRET-melting strategy: it can be performed in 96- microwell plates and allows fast temperature controlling. Additionally, only one dual-labeled strand (F21T) is used in the new assay, which is cutting down expenses. Thanks to the high affinity of PhenDC3 to all G4 topologies, all G4-forming competitors can theoretically be characterized by FRET-MC, no matter their molecularity. Nonetheless, our results show that some G4 competitors with low thermal stability are unfolded before F21T starts to melt, therefore PhenDC3 still stabilizes F21T, producing false negatives. However, the false conclusion is caused by the unavoidable melting procedure in which FRET-MC is based on.

In the second work, we converted the thermal FRET assay into an isothermal version, to solve the false negative generated during the heating process. The isothermal FRET (iso-FRET) assay is based on a duplex-quadruplex competition and (again) the well-characterized bis-quinolinium G4 ligand, PhenDC3. The competitive binding of PhenDC3 between an unlabeled competitor and the BHQ1-labeled 37 mer G4 strand (FRET acceptor: 37Q) drives the results of the isothermal FRET assay, and a FAM-labeled partially complementary strand to 37Q (FRET donor: F22) is then added to report on the structure of 37Q. The FRET occurs and fluorescence is quenched only when 37Q and F22 form a duplex, implying that PhenDC3 binds to the G4 competitor. Differently, if PhenDC3 binds to 37Q, duplex generation is hindered and a high fluorescence of F22 is observed. Although iso-FRET solved the issue associated to the detection of G4s with low T_m , it rises a new problem: some G4 forming competitors can hybridize to the C-rich F22 sequence, preventing the formation of F22-37Q duplex, leading to false negative results. We defined a parameter to estimate the complementarity level between the competitor and F22, to exclude the competitors which have a high probability to hybridize to F22. These two FRET-based G4 methods reiterate that the cross validation by different biophysical methods is necessary for the experimental characterization of PQS *in vitro*.

The last part of my thesis was dedicated to understanding quadruplex-hairpin competition in G4-prone sequences containing runs of consecutive cytosines. Our global mapping results have shown that about 10% of PQS in the human genome (hg19) contains at least three continuous cytosines, which have the possibility to form GC base pairs and disrupt the G4 structure, especially in a potassium-deficient environment. Recent studies suggested that the presence or absence of G4 structures, controlled by potassium-deficiency and potassium / sodium unbalance, can regulate oncogene transcription [245] and pre-mRNA alternative splicing [246]. However, most *in vitro* studies focused on characterizing G4 formation and stability are performed in monocationic buffers (potassium or sodium), which are far apart from the real intracellular

conditions. To investigate the structures formed by sequences rich both in cytosines and guanines in mixed potassium / sodium buffers, we took as a starting point the model sequence 25CEB (PDB: 2LPW), characterized by an extremely long central loop. For the purpose of this study, the central loop bases were replaced by cytosines one by one, and the structures of the mutated sequences were characterized in buffers containing different potassium / sodium ratio, using some of the methods described in the previous §. The effect produced by the presence of cytosines on the G4-hairpin equilibrium and G4 topologies was then analyzed in depth.

Chapter II. FRET melting competition assay

Driven by the classical FRET-melting assay used to characterize G4 formations and G4-ligand complex generation, we first reported on a high-throughput fluorescence-based FRET melting competition assay, the so-called FRET-MC. This assay is based on the competitive binding of a selective G4 ligand (PhenDC3) between a dual-labeled G4 reporter (FAM-Tel21-TAMRA, abbreviated to F21T) and an unknown competitor present in a large excess. In these conditions, a G4-forming competitor can act as a decoy and quantitatively trap PhenDC3 present in solution. As a consequence, a negligible amount of PhenDC3 is available to stabilize F21T and its T_m remains low. In contrast, if the competitor is unable to trap PhenDC3 (non G4 competitor), the latter binds to and stabilizes F21T, increasing its T_m value. Hence, the difference in T_m measured between a system containing PhenDC3 and F21T alone and in the presence of competitor defines if the competitor folds into G4 structure or not.

Below, we report a representative experiment to show the result obtained by FRET-MC for a selection of sequences taken from a plant genome (part of an article added in annex in which I contributed to the biophysical part). The graph bar represented in **Figure 13** shows the ΔT_m values measured for F21T in the presence of PhenDC3 alone or in the presence of PhenDC3 and competitors. Low ΔT_m are obtained in the presence of G4-forming competitor, while high ΔT_m relates to non G4 competitors. Two stable G4 sequences, Pu24T and cmyc, were chosen as positive controls, while a duplex (ds26) and a single strand (dT26) as negative controls. To quantitate the competition effect, we normalized the ΔT_m into the so-called *S Factor*, which corresponds to the relative PhenDC3 stabilization remaining in the presence of the competitor, and clarify the empirical boundaries of G4 and non G4 competitors [247]: *i*) $S < 0.3$: G4 competitor; *ii*) $0.3 \leq S < 0.6$: unknown; *iii*) $S \geq 0.6$: non-G4 competitor.

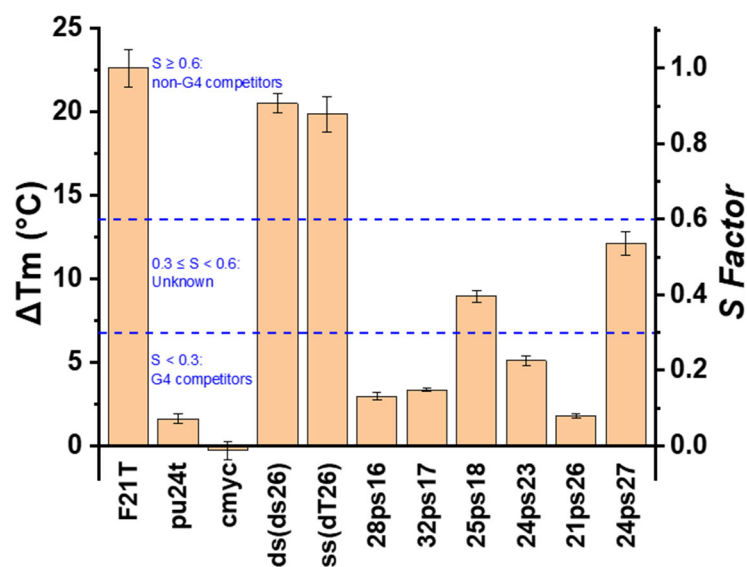


Figure 13 FRET-MC results for the *Pisum sativum* chloroplast sequences. Pu24T and cmyc represent G4-forming positive controls, while ss(dT26) and ds(ds26) correspond to single- and double-stranded negative controls. F21T corresponds to the delta T_m observed in the absence of any competitor ($S = 1$). Samples were measured in 10 mM KCl + 90 mM LiCl, 10 mM LiCaco buffer (pH =7.2). The figure was adapted from [248].

The interaction between the competitor and F21T should be considered. The majority of competitors did not show any complementarity with F21T and ΔT_m was not affected (*i.e.*, RND7, blue & green lines in **Figure 14a**). However, the single-stranded competitor ss1 can hybridize with F21T (**Figure 14b**), leading to reversed F21T melting curves (purple & yellow lines in **Figure 14a**): the highest fluorescence were recorded at low temperature, since ss1 is able to form a duplex with F21T, increasing the distance between the two fluorescent probes located on the two ends of the G4 sequence. Interestingly, the fluorescence of ss1-F21T complex was significantly higher than the one of F21T alone or in the presence of ss1; this might be explained by the original intermolecular attraction of FAM and Tamra [242]. In the F21T-ss1 complex, the rigid ds structure stretched the two termini of F21T extremely far and completely abolished FRET. With the temperature increasing, F21T is released from the rigid duplex, FAM and Tamra are allowed to interact and induce FRET, leading to fluorescence quenching. In other words, high temperature unfolds F21T G4 and increases the distance between its two ends, reducing FRET efficiency and restoring FAM fluorescence to some extent, but not as high as for a duplex.

Although the interaction between the competitor and F21T gives inaccurate result, there is no necessity to check the complementarity degree between F21T and the competitor in FRET-MC assay. Our data evidenced that low complementarity between F21T and the competitor barely affect results, while highly

complementary competitors hybridizing to F21T (*i.e.*, ss1) show totally different melting profiles. Competitor with high complementarity towards F21T can be easily detected by data treatment.

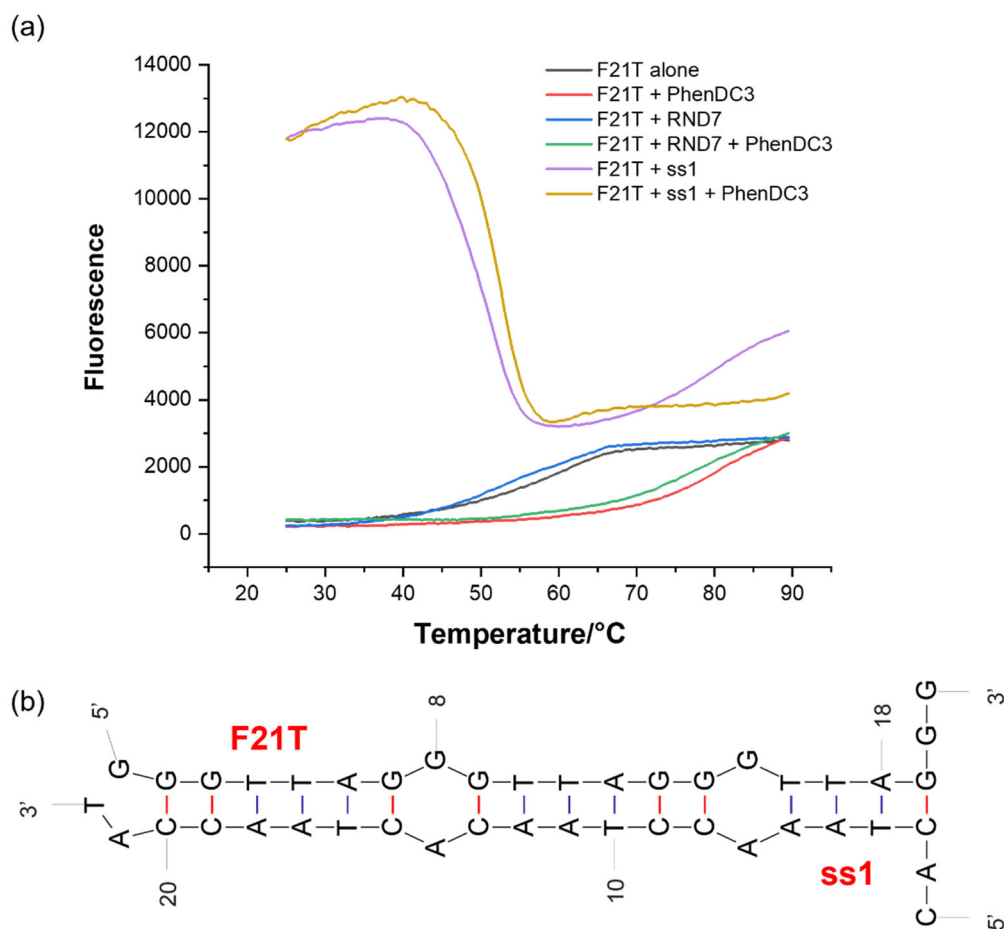


Figure 14 (a) FRET melting curves obtained in the presence of competitors with low complementarity to F21T (*i.e.*, single-stranded RND7 d[GCCTTGC GGAGGCATGCGTCATGCT]) and competitors with high complementarity to F21T (*i.e.*, single-stranded ss1 d[CACTAAACCTAAACCTAACCAT]) in the absence or presence of PhenDC3. Samples were measured in 10 mM KCl, 90 mM LiCl, 10 mM LiCaco buffer (pH = 7.2). (b) Duplex generated by F21T and ss1, the duplex model is simulated by the UNAFold [249].

FRET-MC has been validated with several known sequences and then applied for the *in vitro* characterization of thousands of PQS used in other projects [93, 248, 250] (these articles I co-signed appear in the Annex). These examples can be considered as medium to large scale validations of the FRET-MC protocol. FRET-MC works very well in most cases. Thanks to the excellent selectivity of PhenDC3, we have not met any false positive result so far. As PhenDC3 has high affinity to all G4 topologies and little bias between RNA and DNA G4 structure, FRET-MC can be used to validate all types of G4s in theory, no matter their topology and type of oligonucleotide. However, as the heating process is unavoidable to record T_m values, FRET-MC fails to identify G4-motifs with low thermal stability (*i.e.*, KRAS-22RT, SP-PGQ-3 and TBA) as they appear as

single-stranded at the temperature where F21T unfolds (and one may argue that thermally relatively unstable G4 are *less* likely to be biologically relevant). These false negative results have been shown in the assay validation part [247]. Neither increasing competitor concentration nor extending competitive binding time “save” these weak G4s. Again, there is no perfect assay for G4 characterization. FRET-MC should be reconfirmed by other independent characterizations, especially for samples show negative results. One may also argue that “weak” G4 are unlikely to form in cells, especially when duplex unfolding is required.

ARTICLE

FRET-MC: A fluorescence melting competition assay for studying G4 structures *in vitro*

Yu Luo^{1,2}  | Anton Granzhan¹  | Daniela Verga¹  | Jean-Louis Mergny² 

¹Université Paris Saclay, CNRS UMR9187, INSERM U1196, Institut Curie, Orsay, France

²Laboratoire d'Optique et Biosciences, Ecole Polytechnique, CNRS, Inserm, Institut Polytechnique de Paris, Palaiseau, France

Correspondence

Daniela Verga, Université Paris Saclay, CNRS UMR9187, INSERM U1196, Institut Curie, 91400 Orsay, France.

Jean-Louis Mergny, Laboratoire d'Optique et Biosciences, Ecole Polytechnique, CNRS, Inserm, Institut Polytechnique de Paris, 91128 Palaiseau, France.

Email: jean-louis.mergny@inserm.fr

Funding information

Centre National de la Recherche Scientifique; École Polytechnique, Université Paris-Saclay; Institut Curie; Institut National de la Santé et de la Recherche Médicale

Abstract

G-quadruplexes (G4) play crucial roles in biology, analytical chemistry and nanotechnology. The stability of G4 structures is impacted by the number of G-quartets, the length and positions of loops, flanking motifs, as well as additional structural elements such as bulges, capping base pairs, or triads. Algorithms such as G4Hunter or Quadparser may predict if a given sequence is G4-prone by calculating a quadruplex propensity score; however, experimental validation is still required. We previously demonstrated that this validation is not always straightforward, and that a combination of techniques is often required to unambiguously establish whether a sequence forms a G-quadruplex or not. In this article, we adapted the well-known FRET-melting assay to characterize G4 in batch, where the sequence to be tested is added, as an unlabeled competitor, to a system composed of a dual-labeled probe (F21T) and a specific quadruplex ligand. PhenDC3 was preferred over TMPyP4 because of its better selectivity for G-quadruplexes. In this so-called FRET-MC (melting competition) assay, G4-forming competitors lead to a marked decrease of the ligand-induced stabilization effect (ΔT_m), while non-specific competitors (e.g., single- or double-stranded sequences) have little effect. Sixty-five known sequences with different typical secondary structures were used to validate the assay, which was subsequently employed to assess eight novel sequences that were not previously characterized.

KEYWORDS

DNA structure, FRET-melting, G-quadruplex, G-quartet, UV-melting

1 | INTRODUCTION

G-quadruplexes (G4) are four-stranded nucleic acid structures adopted by G-rich DNA and RNA sequences. G4 result from the stacking of two or more G-quartets (also called G-tetrads), which are formed by four guanine bases interacting through Hoogsteen hydrogen bonds.^[1] G4 structures have been widely used to design biosensors to detect specific small molecules^[2–5] and to control the assembly of supramolecular DNA complexes.^[6–9] G4 structures also exist *in vivo* and play important roles in cells.^[10,11] For example, G4s can be formed at human telomeres and G4 ligands may interfere with telomeric functions, leading to telomere shortening and/or uncapping.^[12–14] G4s are also found in the promoter regions of genes critical in cancer, including *KRAS*, *BCL2*, and

VEGF^[15–17]; *BCL2* plays an essential role in cell survival; *VEGF* is a key angiogenic growth factor which contributes to angiogenesis and tumor progression, and *KRAS* is one of the most frequently mutated oncogenes in many signal transduction pathways, relevant for different types of human carcinomas.^[16]

Given the importance of G4 structures in biology and nanotechnology,^[10] algorithms such as G4Hunter^[18,19] have been developed to predict if a specific sequence is G4-prone. However, for most DNA or RNA motifs, experimental validation is required and, for this purpose, a number of biophysical methods have been developed to characterize G4 structures *in vitro*. Some classical methods are based on the physical properties of G4 structures, such as UV-melting at 295 nm,^[20] nuclear magnetic resonance (NMR),^[21] circular

dichroism (CD) spectroscopy,^[22] isothermal difference spectroscopy (IDS) and thermal difference spectroscopy (TDS).^[23] Fluorescence light-up assays employing dyes such as Thioflavin T,^[24] N-methylmesoporphyrin IX (NMM),^[25] tailor-made dyes^[26] or combinations of dyes^[27,28] are also used to evidence G4 formation. Although there is a wide range of choices to check if a quadruplex is formed or not, this validation is not always straightforward; therefore, a combination of techniques is often required to unambiguously establish whether a sequence folds into a quadruplex or not.^[18]

High-affinity G4 ligands can stabilize a G4 structure and alter its biological functions.^[29] Typical assays used to characterize these ligands are the Fluorescent Intercalator Displacement (FID) assay^[30] and the fluorescence-based Förster Resonance Energy Transfer (FRET)-melting assay.^[31,32] The FRET-melting assay is based on the stabilization induced by a quadruplex ligand, leading to a difference in melting temperature (T_m) between the nucleic acid alone and in presence of this ligand^[33]; in the presence of the latter, the thermal stability of the structure increases in a concentration-dependent manner. This FRET-melting assay has been extensively used to estimate whether a compound is a good quadruplex ligand or not,^[33] despite biases when ranking ligands potency using melting experiments.^[34] More recently, this assay was adapted to assess G4 ligands in near-physiological conditions.^[35]

In this report, instead of testing unknown compounds, we make use of one of the most-characterized G4 ligands, PhenDC3,^[36] to determine if an unknown DNA sequence forms a G-quadruplex structure. PhenDC3 is a high affinity G4 ligand capable of binding to a variety of G4 structures, but with a low affinity for other conformations.^[37,38] We took advantage of these observations to design a novel FRET-melting competition assay, termed FRET-MC, in which the interaction between a fluorescent G4-forming oligonucleotide and PhenDC3 is challenged by the (unlabeled) sequence of interest added in excess. Sixty-five sequences with a known structure were tested to validate this FRET-melting competition assay, which allowed us to determine whether a sequence forms a stable quadruplex or not. Finally, eight novel sequences were used to determine if this assay was accurate: the conclusions reached by this technique were supported by other biophysical methods (CD, IDS, and TDS). Advantages and drawbacks of the FRET-MC method are discussed.

2 | MATERIALS AND METHODS

2.1 | Samples

Oligonucleotides were purchased from Eurogentec, Belgium, as dried samples: unmodified oligonucleotides were purified by RP cartridge while a dual-labeled F21T ((FAM-d((G₃T₂A)₃G₃)-TAMRA) was purified by RP-HPLC. Stock solutions were prepared at 100 μ M strand concentration for the unlabeled oligonucleotide and at 200 μ M strand concentration for F21T in ddH₂O. Fifty micromolar oligonucleotide solutions were annealed in the corresponding buffer, kept at 95 °C for 5 minutes and slowly cooled to room temperature before measurement. The FRET

buffer contains 10 mM KCl, 90 mM LiCl, 10 mM lithium cacodylate, pH 7.2. The K-100 buffer contains 100 mM KCl, 10 mM lithium cacodylate, pH 7.2.

2.2 | FRET-melting competition assay

FRET melting experiments were performed in 96-well plates using a 7900HT RT-PCR instrument (Applied BioSystems). Each well contained competitors either at a single concentration (3 μ M), or at 6 different concentrations, ranging from 0.2 to 3 μ M. 0.2 μ M of the fluorescent oligonucleotide F21T was incubated with or without 0.4 μ M G4 ligand (PhenDC3 for most experiments; TMPyP4 was used in a control experiment shown in Supporting Information) in FRET buffer in a final volume of 25 μ L. The FAM channel was used to collect the fluorescence signal. Samples were kept at 25 °C for 5 minutes, then the temperature was increased by 0.5 °C per minute until 95 °C. Each experimental condition was tested in duplicate on at least two separate plates.

ΔT_m is determined as the difference in F21T T_m with the sample containing competitors in the presence or absence of PhenDC3. The T_m of an oligonucleotide is defined as the temperature at which 50% of the oligonucleotide is unfolded. The most common method to obtain T_m values is approximated as corresponding to half of the height at the normalized melting curve.^[39]

The FRET-melting assay uses a 96-well plate as a sample holder, which allows to process 48 sequences simultaneously. The traditional “midpoint” determination requires a manual analysis.^[40] Given the number of profiles to be analyzed, this process is time consuming as the curves are analyzed one by one. The DoseResp Function in Origin Pro package allows collect T_m in batch. As shown in Figure S1A, Log X_0 in fitting curve is the T_m . This method can only be used in curve (i) (see results) as non-linear fitting would fail if there is no high plateau at row curve. In general, the difference of T_m calculated by these two methods is very small: when using both the “1/2 height” and non-linear fitting methods to calculate T_m of F21T alone, the results are 59.8 °C and 59.4 °C, respectively (Figure S1B). In a few rare cases that should be noted (Figure S1C), curves are irregular, and the minimum of X-axis (usually 25 °C in FRET-melting assay) does not correspond to $Y = 0$ for the normalized curve. In this case, using “1/2 height” is more accurate; T_m calculated by DoseResp always referenced the X value corresponding to $Y = 0.5$, while in some curves the 1/2 height ($Y_x = 95 - Y_x = 25$) at Y-axis is not always 0.5. In the instance, 1/2 height at Y-axis is 0.525, T_m calculated by “1/2 height” is 53.0 °C, to be compared with 51.8 °C by DoseResp.

$$\text{DoseResp Function: } Y = A_1 + \frac{A_2 - A_1}{1 + 10^{(\log x_0 - x)p}}$$

In brief, DoseResp is only truly accurate for curves with appropriate upper and lower baselines, while the “1/2 height” approach can be applied in all cases, provided that normalization is accurate. If possible, we suggest to use the same method to calculate T_m for

experiments performed in parallel, although differences of T_m determined by these two approaches is small in general.

2.3 | UV-melting assay

UV-melting curves of 5 μ M oligonucleotides in FRET buffer were recorded with a Cary 300 (Agilent Technologies, France) spectrophotometer. Heating runs were performed between 10 $^{\circ}$ C and 95 $^{\circ}$ C, the temperature was increased by 0.2 $^{\circ}$ C per minute, and absorbance was recorded at 260 and 295 nm. T_m was determined as the temperature corresponding to half of the height of the normalized melting curve.

2.4 | Circular dichroism

Three micromolar SP-PGQ-1 was kept in 1000 μ L FRET buffer. Three micromolar Oligonucleotides of testing set were kept in 1000 μ L K-100 buffer. CD spectra were recorded on a JASCO J-1500 (France) spectropolarimeter at room temperature, using a scan range of 400 to 230 nm, a scan rate of 200 nm/min and averaging four accumulations.

2.5 | Thermal difference spectra

Three micromolar SP-PGQ-1 was kept in FRET buffer. Three micromolar Oligonucleotides of testing set were kept in 1000 μ L K-100 buffer. Absorbance spectra were recorded on a Cary 300 (Agilent Technologies, France) spectrophotometer at 25 $^{\circ}$ C (scan range: 500–200 nm; scan rate: 600 nm/min; automatic baseline correction). After

recording the first spectrum (folded), temperature was increased to 95 $^{\circ}$ C, and the second UV-absorbance spectrum was recorded after 15 minutes of equilibration at high temperature. TDS corresponds to the arithmetic difference between the initial (folded; 25 $^{\circ}$ C) and second (unfolded; 95 $^{\circ}$ C) spectra.

2.6 | Isothermal difference spectra

Three micromolar Oligonucleotides of testing set were kept in 900 μ L K-100 buffer. Absorbance spectra were recorded on a Cary 300 (Agilent Technologies, France) spectrophotometer at 25 $^{\circ}$ C (scan range: 500–200 nm; scan rate: 600 nm/min; automatic baseline correction). A hundred microliter of 1 M KCl was added after recording the first spectrum, and the second UV-absorbance spectrum was recorded after 15 minutes of equilibration. IDS correspond to the arithmetic difference between the initial (unfolded) and second (folded, thanks to the addition of K^+) spectra, after correction for dilution.

3 | RESULTS AND DISCUSSION

3.1 | Principle of the FRET-MC assay

In a K^+ -containing buffer, F21T forms a stable G4 structure which can be used as a FRET-melting probe thanks to the fluorophores attached at both ends: a 5'-appended donor (Fluorescein) and a 3'-appended acceptor (TAMRA), allowing efficient energy transfer between the donor and acceptor dyes when the oligonucleotide is folded (Figure 1). Thermal unfolding leads to the disruption of the

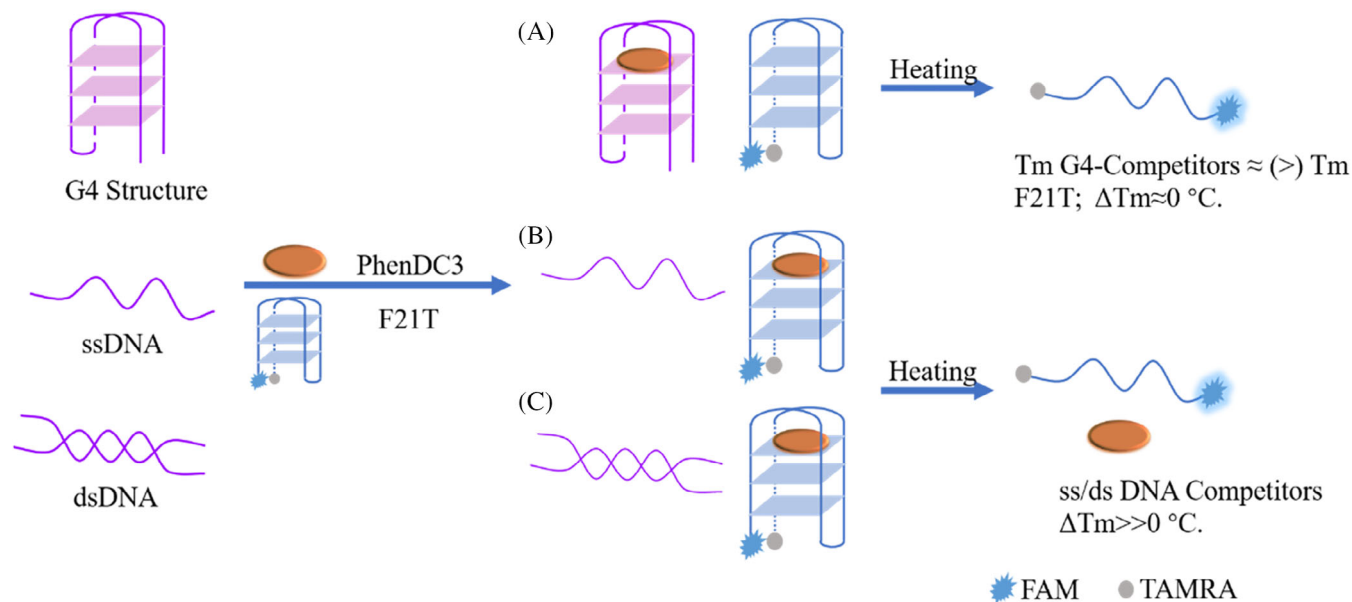


FIGURE 1 Principle of the FRET-MC assay. In panel A, the competitor forms a quadruplex and traps PhenDC3 (shown as an orange oval). In panels B and C, the competitor does not form a quadruplex and has no affinity for PhenDC3, which remains bound to F21T (in light blue). Competitor sequences are shown in purple, PhenDC3 is represented as an orange oval, and F21T is represented in blue

G-quadruplex structure and a decrease in FRET efficiency, as the 5' and 3' ends become distant when the sequence is single-stranded. In the presence of a G4 ligand, the melting temperature (T_m) of F21T increases, as the ligand makes the structure of F21T more stable. F21T is typically used in a FRET-melting assay, in which the specificity of a ligand is tested by adding various specific (G4-forming) and unspecific (duplexes or single-stranded) competitors.^[32] In this report, we are radically changing our viewpoint: rather than testing ligands of unknown specificity against known competitors, we challenge of the best-characterized G4-ligand, PhenDC3, with a variety of oligonucleotide competitors. Adding a large excess of an unlabeled oligonucleotide may lead to two possible scenarios:

- i. The competitor is unable to trap the quadruplex ligand. In this case, T_m of the (F21T + PhenDC3) system is not affected by the competing oligonucleotide (in other words, ΔT_m remains high). This is the expected outcome for a single-strand or a DNA or RNA duplex.
- ii. If the unlabeled competitor has a high affinity for PhenDC3, it will sequester a significant fraction of the compound, which will be no longer available for F21T stabilization, leading to a decrease in T_m . In this case, if the competition is very efficient, the T_m should fall back close to the value obtained without quadruplex ligand, meaning for F21T alone (in other words, $\Delta T_m \approx 0$).

To proceed, we selected a variety of competitors for which the structure was previously investigated and characterized.^[41] This collection of over 60 sequences includes a variety of quadruplex-forming motifs (with various topologies) as well as single- and double-stranded DNAs (sequences shown in Table S1).

3.2 | Validation of the FRET-melting competition assay with a set of 65 sequences

A trivial, but important, control was first performed by checking that the competitors do not directly interact with F21T. To that aim, T_m of F21T was measured alone or in the presence of each competitor, in the absence of PhenDC3. As expected, most tested sequences had negligible, if any, effect on F21T melting (Figure S2; normalized FRET-melting curves are shown in Figure S3). A few motifs (46AG, T95-2T, T2B-1, AT11, LWDLN1, AND1, RND3, RND6 and AT26) led to a significant decrease in T_m ($\Delta T_m > 5^\circ\text{C}$).

We next investigated the impact of the competitors on the stabilization effect (ΔT_m) induced by PhenDC3 on F21T. Figure 2 presents examples of FRET-melting profiles for F21T alone (Figure 2A), F21T in competition with a single-strand (Figure 2B), a stable G4 (Figure 2C) or a duplex DNA (Figure 2D), respectively.

In the absence of any competitor, 2 eqv. of PhenDC3 induces a ΔT_m of 23.4°C , in agreement with previous results.^[42] As expected, none of the competitors induced a further significant increase in T_m , as compared to F21T + PhenDC3. Figure 3 summarizes the ΔT_m results obtained for F21T in the presence of

PhenDC3 and in the presence or absence of oligonucleotide competitor (normalized FRET-melting curves shown in Figure S3). Many, but not all, of the sequences known to form G4 structures led to a significant drop in ΔT_m values (upper half of the figure), showing that they acted as efficient competitors. In contrast, single-stranded and duplex DNAs had little impact, as ΔT_m values remained high, close to the value found for F21T + PhenDC3 with no competitor. To quantitate this competition effect, we defined the *S Factor*, as originally described in Reference [43], which corresponds to the relative PhenDC3 stabilization remaining in the presence of the competitor:

$$S\text{Factor} = \frac{\Delta T_m \text{ of F21T with competitors}}{\Delta T_m \text{ of F21T alone}}$$

Based on *S Factor*, the competitors can be divided into two categories: (i) ineffective competitors, for which *S* remains ≈ 1 , meaning that competition is nearly completely unproductive, as expected for a structure for which PhenDC3 has no affinity, and conversely (ii) potent competitors (i.e., stable quadruplexes) would give a *S Factor* close to 0 (Figure 3).

As expected, the majority of G4-forming sequences led to *S* values close to 0. However, several known G4 structures (UpsB-Q3, SP-PGQ-1, KRAS-22RT, SP-PGQ-3, TBA, Bm-U16, Bom17 and LWDLN3) were not efficient competitors ($S > 0.6$). In order to understand these results, we performed UV-melting experiments for these 8 sequences and collected T_m values in Table S2 (detailed UV-melting curves are shown in Figure S4). The thermal stability of all these sequences (except for SP-PGQ-1) was relatively low, indicating they form unstable G4 structures, which are likely to be unfolded in the temperature range where F21T starts to melt: they are then “seen” as non-specific single-strands rather than true G-quadruplexes. SP-PGQ-1 behaved differently: UV-absorbance at 295 nm of a quadruplex should decrease upon heating due to the unfolding of the quadruplex structure. Although SP-PGQ-1 was reported to form a hybrid G4 structure,^[44] our results show an unexpected increase in absorbance at 295 nm upon UV-melting, incompatible with the unfolding of a quadruplex, and rather suggesting the formation of another structure (e.g., a mismatched duplex) at low temperatures. CD spectroscopy and TDS confirmed this hypothesis (Figure S5). In contrast, the T_m of 10 different G4 sequences acting as effective competitors ($S < 0.3$ for 46AG, Bcl2Mid, 25TGA, Chl, LTR-III, c-kit-T12T2, Pu24T, c-kit87up, VEGF, and T95-2T) were always higher than the T_m of the false negative sequences, as shown in Table S2 (UV-melting curves showed in Figure S4). Although some of the T_m are still lower than the T_m of F21T, their presence in large excess (15-fold molar excess as compared to F21T) may compensate for a partial denaturation.

We then investigated whether one could substitute PhenDC3 by another G4 ligand, TMPyP4, a cationic porphyrin which also has a high affinity for G-quadruplexes, but is much less selective. Figure S6 presents ΔT_m and *S* values for $0.4\ \mu\text{M}$ TMPyP4 on $0.2\ \mu\text{M}$ F21T, alone or in the presence of various competitors at $3\ \mu\text{M}$ strand

concentration. Experiments were done in a buffer identical to the one used for PhenDC3. We tested five representative sequences for each structural type considered here (25 different competitors in total). Although parallel quadruplexes were the most efficient competitors, some single-strands and most duplexes were also competing, with S value around 0.5, lower than the S values found for most anti-parallel and two hybrid quadruplexes with TMPyP4. This experiment demonstrates that, for this method to be reliable, a truly G4-specific ligand without preferential binding to any G4 topology must be chosen. While other compounds than PhenDC3 may fit the bill, moderately selective compounds will not.

A critical factor for the competition efficiency (measured by the S value) in the FRET-MC assay should be the affinity and number of binding sites present on the competitor sequence for the quadruplex ligand (PhenDC3 in all further experiments). We wanted to investigate if the thermal stability of the structure itself would also contribute. All

duplex sequences tested here had a significantly higher T_m value than F21T (as shown in Table S2); they were still poor competitors, even when added in large excess, as PhenDC3 is unable to bind to duplexes.

For the quadruplexes, in order to investigate how S value correlates with T_m , we tried a range of six different competitor concentrations (1x to 30x molar excess, as compared to F21T) for nine different G4-forming sequences. As shown in Figure S7, the T_m for each quadruplex (detailed UV-melting curves are shown in Figure S4) was experimentally determined under identical conditions, and is shown in red below; sequences were ranked from left (lowest T_m) to right (highest T_m). Overall, there is indeed a correlation between T_m and S values: the competitors with a high thermal stability lead to low S values when the competitor is not in huge excess. Since S values are low for nearly all sequences at 3 or 6 μM , independently of T_m , we did not consider these two concentrations in Figure S8, which provides a

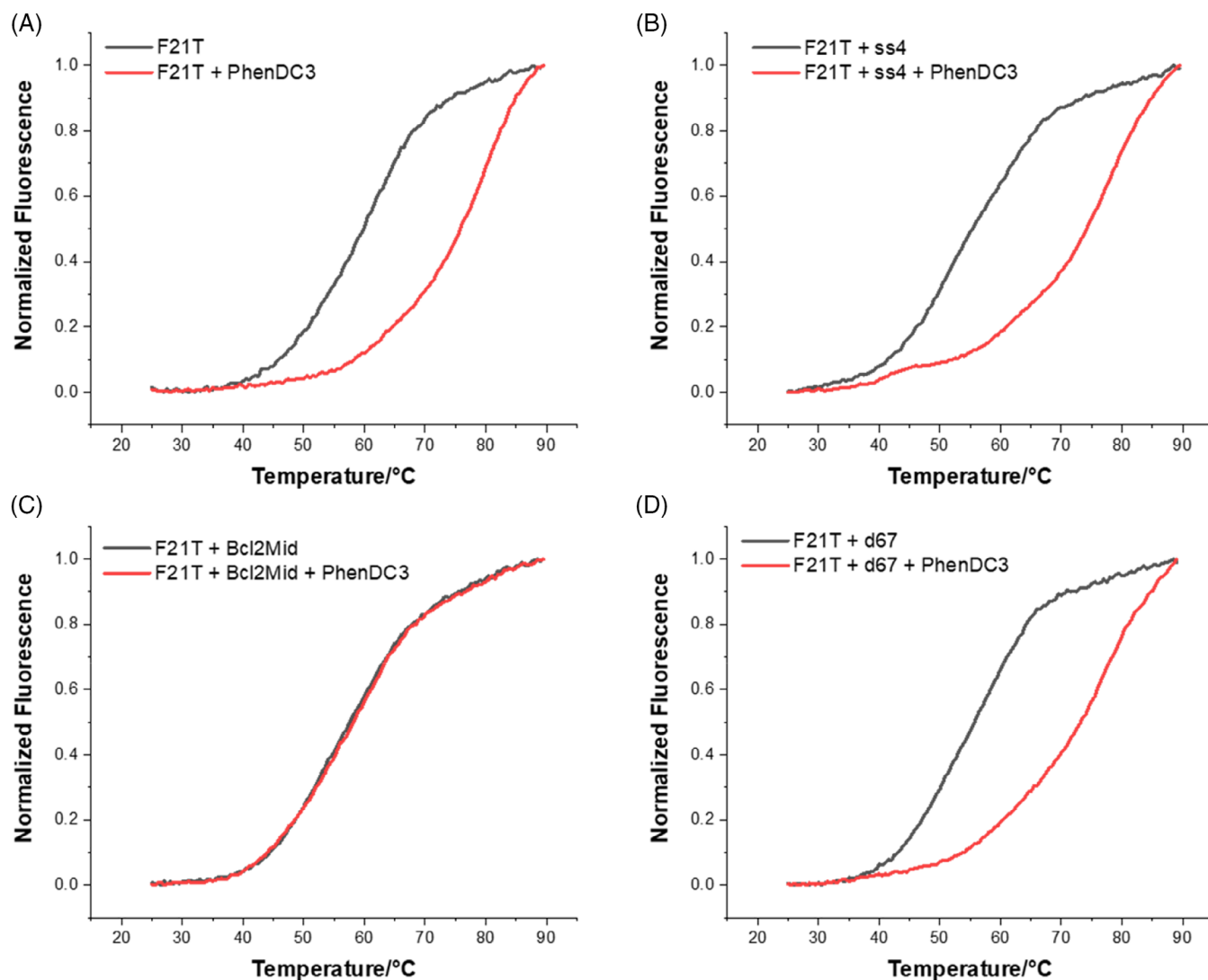


FIGURE 2 FRET-melting profiles of 0.2 μM F21T alone or in the presence of 3 μM competitors, and with (red) or without (black) 0.4 μM PhenDC3. A, F21T alone and in the presence of PhenDC3, and with the addition of the competitor: B, single-strand, C, quadruplex DNA and D, duplex DNA. Samples were annealed and measured in 10 mM KCl, 90 mM LiCl, 10 mM lithium cacodylate pH 7.2 buffer

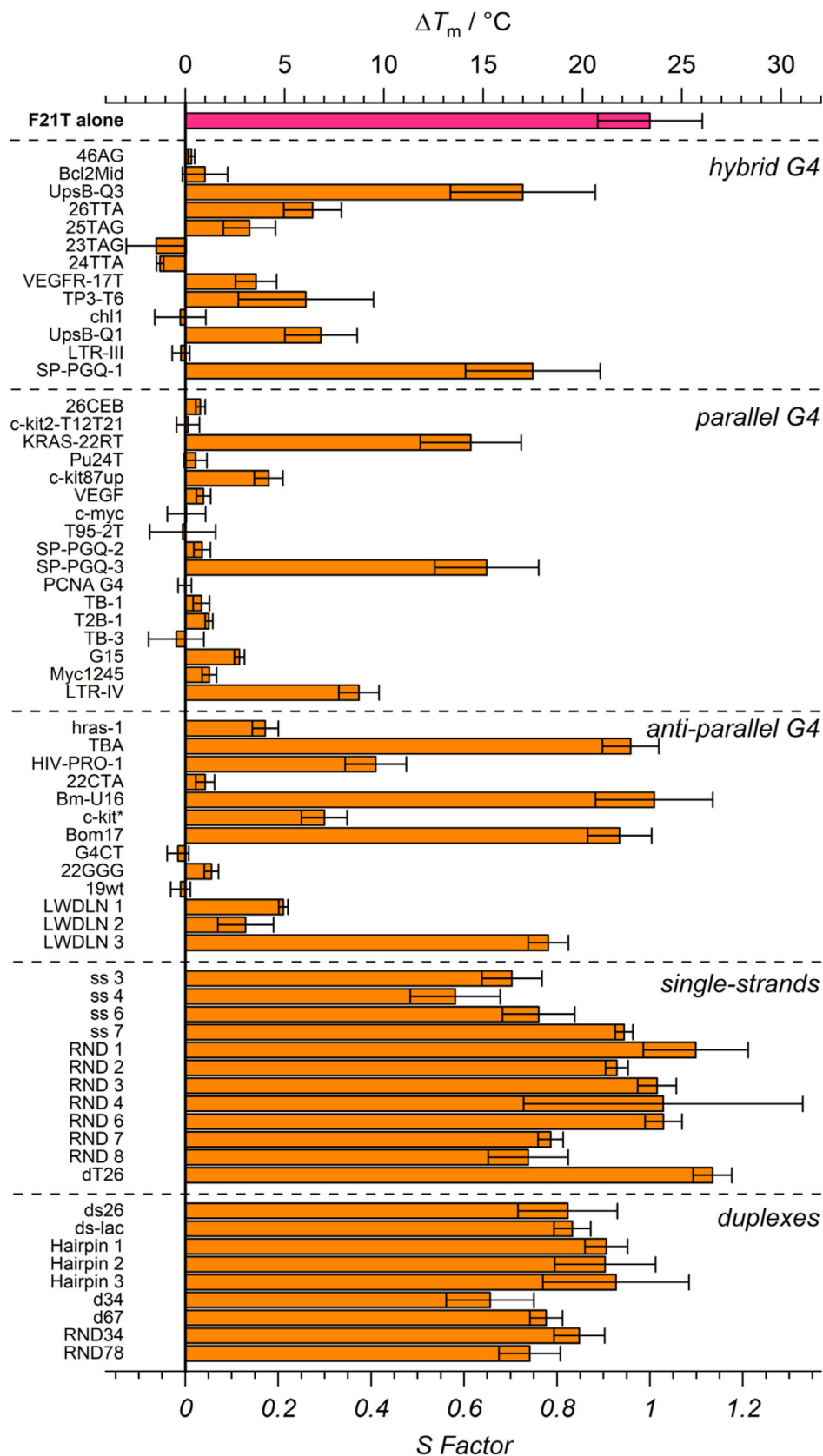


FIGURE 3 ΔT_m induced by $0.4 \mu\text{M}$ PhenDC3 on $0.2 \mu\text{M}$ F21T, alone or in the presence of $3 \mu\text{M}$ competitors. The S Factor (bottom X-axis) provides a normalized value. Samples were annealed and measured in 10 mM KCl, 90 mM LiCl, 10 mM lithium cacodylate pH 7.2 buffer

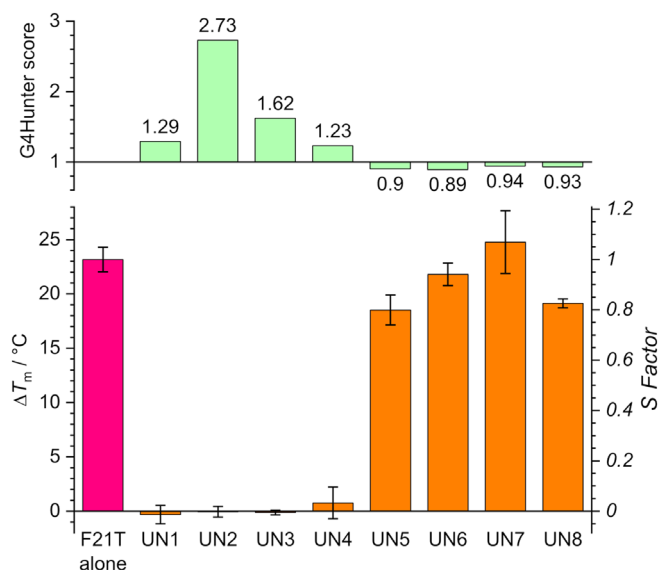


FIGURE 4 Testing eight different sequences (UN1-UN8) with the FRET-MC assay. The G4Hunter score for each competitor is indicated on the upper part of the figure (green bars). The ΔT_m induced by 0.4 μ M PhenDC3 on 0.2 μ M F21T, alone or in the presence of 3 μ M competitors (UN1-8) is plotted on the lower part of the graph. The *S Factor* (right Y-axis) provides a normalized value. Samples were annealed and measured in 10 mM KCl, 90 mM LiCl, 10 mM lithium cacodylate pH 7.2 buffer

different view/representation of these *S* values, in which sequences are clustered in three categories according to T_m . The most striking difference is found between the high stability group and the others. Table S3 summarizes average *S* values (from 1x to 10x molar excess) for these nine sequences. Generally, average *S* values decreased with increasing T_m but the correlation is far from perfect. The average *S* values for high stability sequences are 0.25 or below, while *S* values for the others are above 0.35.

3.3 | Experimental validation on novel sequences

To test this method on a set of novel sequences, we used eight genomic DNA sequences of unknown structure. These motifs are found close to the transcription start site of different human promoters, and their biological relevance is currently being investigated. According to G4Hunter analysis,^[18] four of them (UN1-4) were likely to form a quadruplex structure, as their G4Hunter score was above 1.2, while the remaining four (UN5-8) were unlikely to adopt a G4 conformation, with a G4Hunter score < 1.0. Detailed sequences with location information (Human hg19) and G4Hunter scores are shown in Table S4.

The ΔT_m values and *S Factors* for the tested sequences are shown on Figure 4. The first four oligonucleotides (UN1-4) with high G4-hunter scores gave low *S Factor* values (<0.1), consistent with G4 formation. In contrast, the remaining four samples (UN5-8) were poor competitors (*S* > 0.8) and unlikely to form quadruplexes, in agreement with their low G4Hunter scores. These conclusions were confirmed by three independent techniques (CD, IDS, and TDS) measured in

100 mM KCl (Figure S9), all giving results consistent with the formation of G4 structures by UN1-4 and the absence of these structures in the case of UN5-8.

4 | CONCLUSION

The FRET-MC assay described here is a fast and inexpensive characterization method to determine if a sequence is forming a stable quadruplex or not. It offers several advantages:

- It is relatively *inexpensive*: while the F21T dual-labeled oligonucleotide is relatively expensive, minimal amounts are used for each point as the volume is reduced (25 μ L) and its concentration is only 0.2 μ M, and can even be further reduced if necessary, provided a sensitive RT-PCR instrument is available. Conversely, the assayed sequences are unmodified oligonucleotides and do not require extensive purification; only 25 picomoles are needed per point.
- It is *fast*: FRET-melting takes 1 to 2 hours while UV-melting requires several hours (14 hours with the temperature gradient used here) for each experiment.
- It is *simple* to set up: all reagents are commercially available (including PhenDC3) and the FRET melting assay is now routinely used in a number of labs.
- It allows testing *multiple samples in parallel*. While a classical UV spectrophotometer can only read up to 6 or 9 samples, FRET-melting uses a 96-well plate as a sample holder, and it is able to process 48 sequences in duplicate. It may even be transposed to 384-well format.

At the same time, this method has several limitations:

- The main assumptions for this technique are that PhenDC3 (i) indiscriminately binds to all G-quadruplexes, and (ii) does not bind to other structures. In other words, for this method to work, we need a perfect, general G4 ligand with high structure specificity. Previously published studies^[30,36] have shown that PhenDC3 does indeed bind to all G4 tested so far, and has excellent specificity. We cannot exclude, however, that PhenDC3 would also recognize other unusual motifs such as G-triplexes.^[45] Additional experiments are therefore required to reach a clear conclusion for a given sequence: as previously stated,^[18] we advocate the use of several independent techniques to assess G4 formation. On the other hand, a high-affinity but poorly selective quadruplex ligand such as TMPyP4 proved to be ineffective for this application.
- The sequence to be tested should not be complementary to F21T, as it would interfere with the structure of the fluorescent probe. C-rich sequences, and especially repetitions of the CCCTAA hexanucleotide motif should be avoided. The training set chosen for this study did not involve any i-motif sequence, resulting from the folding of C-rich oligonucleotides. These sequences would then be partially complementary to the fluorescent G-rich oligonucleotide F21T: duplex formation would “kill” the assay by interfering

with G4 formation. On the other hand, the method itself may later be adapted to the analysis of i-motif sequences. This assay would involve a “complementary” sequence for i-motif formation, in which the fluorescently labeled oligonucleotide is C-rich, not G-rich, and forms an i-motif itself. But this would require a “perfect” i-motif ligand as well, that is, a compound that would bind reasonably well to all i-motifs while having no affinity for any other structure. The i-motif ligands we have tested do not meet these criteria, and cannot be considered as equivalent to PhenDC3 for this purpose.

- The main limitation of this FRET-melting competition assay is its inability to detect unstable quadruplexes which behave as single-strands (Figure 3) at the temperature where F21T starts to melt. This assay should therefore be employed to identify moderately to highly stable G4 structures. A possible way to circumvent this limitation would be to replace or complement F21T by a quadruplex probe with lower stability.^[46] A two-quartet quadruplex such as the thrombin binding aptamer (TBA) could be proposed, keeping in mind that these G4 are often weaker binders for G4 ligands such as PhenDC3. In any case, unstable quadruplexes are less likely to be biologically relevant^[47] and unlikely to be identified by genome-wide methods such as G4-seq^[48] as extension is performed at 60 °C during Illumina sequencing.
- Finally, the *S* value cannot be used as a proxy for the (thermal) stability of the tested quadruplex. While there is some correlation between thermal stability and competition efficiency (stable G4 tend to give lower *S* values), other factors contribute to the competition efficiency, such as the affinity of the PhenDC3 ligand for this topology, and the number of binding sites available.

Overall, despite the shortcomings listed above, the FRET-melting competition assay should constitute an interesting addition to the *in vitro* “G4 characterization toolbox.”

ACKNOWLEDGMENTS

We thank both reviewers for excellent suggestions, Laurent Lacroix (ENS, Paris) for helpful discussions, and Corinne Landras Guetta and Marie-Paule Teulade-Fichou (Institut Curie, Orsay) for a sample of PhenDC3. This manuscript is dedicated to the memory of Prof. Michael J. Waring, with whom Jean-Louis Mergny had interesting lively discussions about DNA ligands during his sabbatical in France.

CONFLICT OF INTEREST

The authors declare no competing interests.

DATA AVAILABILITY STATEMENT

Raw data / melting profiles may be downloaded at <https://data.mendeley.com/datasets/gspc9r73r5/1>.

ORCID

Yu Luo  <https://orcid.org/0000-0003-0614-6150>

Anton Granzhan  <https://orcid.org/0000-0002-0424-0461>

Daniela Verga  <https://orcid.org/0000-0002-7555-6033>

Jean-Louis Mergny  <https://orcid.org/0000-0003-3043-8401>

REFERENCES

- [1] M. Gellert, M. N. Lipsett, D. R. Davies, *Proc. Natl. Acad. Sci. U. S. A.* **1962**, *48*, 2013.
- [2] Y. Guo, L. Xu, S. Hong, Q. Sun, W. Yao, R. Pei, *Analyst* **2016**, *141*, 6481.
- [3] H. Z. He, D. S. Chan, C. H. Leung, D. L. Ma, *Nucleic Acids Res.* **2013**, *41*, 4345.
- [4] J. Ren, T. Wang, E. Wang, J. Wang, *Analyst* **2015**, *140*, 2556.
- [5] B. Ruttkay-Nedecky, J. Kudr, L. Nejdil, D. Maskova, R. Kizek, V. Adam, *Molecules* **2013**, *18*, 14760–79.
- [6] O. Lustgarten, R. Carmieli, L. Motiei, D. Margulies, *Angew. Chem. Int. Ed. Engl.* **2019**, *58*, 184.
- [7] J. L. Mergny, D. Sen, *Chem. Rev.* **2020**, *120*, 11698.
- [8] L. Stefan, D. Monchaud, *Nat. Rev. Chem.* **2019**, *3*, 650.
- [9] Y. Cao, S. Gao, Y. Yan, M. F. Bruist, B. Wang, X. Guo, *Nucleic Acids Res.* **2017**, *45*, 26.
- [10] J. Spiegel, S. Adhikari, S. Balasubramanian, *Trends Chem.* **2020**, *2*, 123.
- [11] D. Varshney, J. Spiegel, K. Zyner, D. Tannahill, S. Balasubramanian, *Nat. Rev. Mol. Cell Biol.* **2020**, *21*, 459.
- [12] A. M. Burger, F. Dai, C. M. Schultes, A. P. Reszka, M. J. Moore, J. A. Double, S. Neidle, *Cancer Res.* **2005**, *65*, 1489.
- [13] M. Read, R. J. Harrison, B. Romagnoli, F. A. Tanious, S. H. Gowan, A. P. Reszka, W. D. Wilson, L. R. Kelland, S. Neidle, *Proc. Natl. Acad. Sci. U. S. A.* **2001**, *98*, 4844.
- [14] J. F. Riou, L. Guittat, P. Mailliet, A. Laoui, E. Renou, O. Petitgenet, F. Mégnin-Chanet, C. Hélène, J. L. Mergny, *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 2672.
- [15] T. A. Brooks, S. Kendrick, L. Hurley, *FEBS J.* **2010**, *277*, 3459.
- [16] J. Jana, S. Mondal, P. Bhattacharjee, P. Sengupta, T. Roychowdhury, P. Saha, P. Kundu, S. Chatterjee, *Sci. Rep.* **2017**, *7*, 40706.
- [17] J. Marqueville, C. Robert, O. Lagrabette, M. Wahid, A. Bourdoncle, L. E. Xodo, J. L. Mergny, G. F. Salgado, *Nucleic Acids Res.* **2020**, *48*, 9336.
- [18] A. Bedrat, L. Lacroix, J. L. Mergny, *Nucleic Acids Res.* **2016**, *44*, 1746.
- [19] V. Brazda, J. Kolomaznik, J. Lysek, M. Bartas, M. Fojta, J. Stastny, J. L. Mergny, *Bioinformatics* **2019**, *35*, 3493.
- [20] J. J. Alba, A. Sadurni, R. Gargallo, *Crit. Rev. Anal. Chem.* **2016**, *46*, 443.
- [21] M. Adrian, B. Heddi, A. T. Phan, *Methods* **2012**, *57*, 11.
- [22] J. Kyrp, I. Kejnovska, D. Renciuik, M. Vorlickova, *Nucleic Acids Res.* **2009**, *37*, 1713.
- [23] J. L. Mergny, J. Li, L. Lacroix, S. Amrane, J. B. Chaires, *Nucleic Acids Res.* **2005**, *33*, e138.
- [24] A. Renaud de la Faverie, A. Guedin, A. Bedrat, L. A. Yatsunyk, J. L. Mergny, *Nucleic Acids Res.* **2014**, *42*, e65.
- [25] N. C. Sabharwal, V. Savikhin, J. R. Turek-Herman, J. M. Nicoludis, V. A. Szalai, L. A. Yatsunyk, *FEBS J.* **2014**, *281*, 1726.
- [26] X. Xie, A. Renvoisé, A. Granzhan, M.-P. Teulade-Fichou, *New J. Chem.* **2015**, *39*, 5931.
- [27] A. Krieg, J. Calvert, J. Sanoica, E. Cullum, R. Tipanna, S. Myong, *Nucleic Acids Res.* **2015**, *43*, 7961.
- [28] M. Zuffo, X. Xie, A. Granzhan, *Chem. Eur. J.* **2019**, *25*, 1812.
- [29] M. P. O'Hagan, J. C. Morales, M. C. Galan, *Eur. J. Org. Chem.* **2019**, *2019*, 4995.
- [30] P. L. Tran, E. Lary, F. Hamon, M. P. Teulade-Fichou, J. L. Mergny, *Biochimie* **2011**, *93*, 1288.
- [31] A. De Cian, L. Guittat, M. Kaiser, B. Sacca, S. Amrane, A. Bourdoncle, P. Alberti, M. P. Teulade-Fichou, L. Lacroix, J. L. Mergny, *Methods* **2007**, *42*, 183.

- [32] J.-L. Mergny, J.-C. Maurizot, *ChemBioChem* **2001**, *2*, 124.
- [33] D. Renciuik, J. Zhou, L. Beaurepaire, A. Guedin, A. Bourdoncle, J. L. Mergny, *Methods* **2012**, *57*, 122.
- [34] A. Marchand, F. Rosu, R. Zenobi, V. Gabelica, *J. Am. Chem. Soc.* **2018**, *140*, 12553–12565.
- [35] R. K. Morgan, A. M. Psaras, Q. Lassiter, K. Raymer, T. A. Brooks, *Biochim. Biophys. Acta Gene Regul. Mech.* **1863**, 2020, 194478.
- [36] A. De Cian, E. DeLemos, J.-L. Mergny, M.-P. Teulade-Fichou, D. Monchaud, *J. Am. Chem. Soc.* **2007**, *129*, 1856.
- [37] A. Marchand, A. Granzhan, K. Iida, Y. Tsushima, Y. Ma, K. Nagasawa, M. P. Teulade-Fichou, V. Gabelica, *J. Am. Chem. Soc.* **2015**, *137*, 750.
- [38] E. Ruggiero, S. N. Richter, *Nucleic Acids Res.* **2018**, *46*, 3270.
- [39] J.-L. Mergny, L. Lacroix, *Curr. Protoc. Nucleic Acid Chem.* **2009**, *37*, 17.1.1–17.1.15.
- [40] J.-L. Mergny, L. Lacroix, *Oligonucleotides* **2003**, *13*, 515.
- [41] M. Zuffo, A. Gandolfini, B. Heddi, A. Granzhan, *Nucleic Acids Res.* **2020**, *48*, e61.
- [42] N. M. Gueddouda, M. R. Hurtado, S. Moreau, L. Ronga, R. N. Das, S. Savrimoutou, S. Rubio, A. Marchand, O. Mendoza, M. Marchivie, L. Elmi, A. Chansavang, V. Desplat, V. Gabelica, A. Bourdoncle, J. L. Mergny, J. Guillon, *ChemMedChem* **2017**, *12*, 146.
- [43] D. Monchaud, C. Allain, H. Bertrand, N. Smargiasso, F. Rosu, V. Gabelica, A. De Cian, J. L. Mergny, M. P. Teulade-Fichou, *Biochimie* **2008**, *90*, 1207.
- [44] S. K. Mishra, N. Jain, U. Shankar, A. Tawani, T. K. Sharma, A. Kumar, *Sci. Rep.* **2019**, *9*, 1791.
- [45] L. Bonnat, M. Dautriche, T. Saidi, J. Revol-Cavalier, J. Dejeu, E. Defrancq, T. Lavergne, *Org. Biomol. Chem.* **2019**, *17*, 8726.
- [46] A. De Rache, J. L. Mergny, *Biochimie* **2015**, *115*, 194.
- [47] A. Piazza, M. Adrian, F. Samazan, B. Heddi, F. Hamon, A. Serero, J. Lopes, M. P. Teulade-Fichou, A. T. Phan, A. Nicolas, *EMBO J.* **2015**, *34*, 1718.
- [48] V. S. Chambers, G. Marsico, J. M. Boutell, M. Di Antonio, G. P. Smith, S. Balasubramanian, *Nat. Biotechnol.* **2015**, *33*, 877.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Luo Y, Granzhan A, Verga D, Mergny J-L. FRET-MC: A fluorescence melting competition assay for studying G4 structures in vitro. *Biopolymers*. 2021; 112:e23415. <https://doi.org/10.1002/bip.23415>

FRET-MC: a fluorescence melting competition assay for studying G4 structures *in vitro*

Yu Luo^{1,2}, Anton Granzhan¹, Daniela Verga^{1*} & Jean-Louis Mergny^{2*}

1. Université Paris Saclay, CNRS UMR9187, INSERM U1196, Institut Curie, 91400 Orsay, France.
2. Laboratoire d'Optique et Biosciences, Ecole Polytechnique, CNRS, Inserm, Institut Polytechnique de Paris, 91128 Palaiseau, France.

* Authors to whom correspondence may be addressed: daniela.verga@curie.fr; jean-louis.mergny@inserm.fr

Supplementary information

Tables S1-S4

Figures S1-S9

Table S1. Training set of DNA sequences.

Name	Sequence (5'-3')	Reported conformation	PDB entry
F21T	FAM-GGGTTAGGGTTAGGGTTAGGG-TAMRA	-	-
46AG	AGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTA GGGTTAGGG	hybrid G4	-
Bcl2Mid	GGGCGCGGGAGGAATTGGGCGGG	hybrid G4	2F8U
UpsB-Q3	CAGGGTTAAGGGTATACATTTAGGGGTTAGGGTT	hybrid G4	-
26TTA	TTAGGGTTAGGGTTAGGGTTAGGGTT	hybrid G4	2JPZ
25TGA	TAGGGTTAGGGTTAGGGTTAGGGTT	hybrid G4	2JSL
23TAG	TAGGGTTAGGGTTAGGGTTAGGG	hybrid G4	2JSK
24TTA	TTAGGGTTAGGGTTAGGGTTAGGG	hybrid G4	2JSL
VEGFR-17T	GGGTACCCGGGTGAGGTGCGGGGT	hybrid G4	5ZEV
TP3-T6	TGGGGTCCGAGGCGGGGCTTGGG	hybrid G4	6AC7
chl1	GGGTGGGGAAGGGGTGGGT	hybrid G4	2KPR
UpsB-Q1	CAGGGTTAAGGGTATAACTTTAGGGGTTAGGGTT	hybrid G4	5MTA
LTR-III	GGGAGGCGTGGCCTGGGCGGGACTGGGG	hybrid G4	6H1K
SP-PGQ-1	GGGCAACTTGGCTGGGGTCTAGTTCCACGGGACGGG	hybrid G4	-
26CEB	AAGGGTGGGTGTAAGTGTGGGTGGGT	parallel G4	2LPW
c-kit2- T12T21	CGGGCGGGCGCTAGGGAGGGT	parallel G4	2KYP
KRAS-22RT	AGGGCGGTGTGGGAATAGGGAA	parallel G4	5I2V
Pu24T	TGAGGGTGGTGAGGGTGGGGAAGG	parallel G4	2A5P
c-kit87up	AGGGAGGGCGCTGGGAGGAGGG	parallel G4	2O3M
VEGF	CGGGCGGGCCTTGGGCGGGGT	parallel G4	2M27
c-myc	TGAGGGTGGGTAGGGTGGGTAA	parallel G4	1XAV
T95-2T	TTGGGTGGGTGGGTGGGT	parallel G4	2LK7
SP-PGQ-2	GGGCTAGTGGGGGAGGGGG	parallel G4	-
SP-PGQ-3	GGGCTAATAGGGAGAGCAGGGACGGGG	parallel G4	-
PCNA G4	CAGGGCGACGGGGGCGGGGCGGGGCG	parallel G4	-
TB-1	TTGTGGTGGGTGGGTGGGT	parallel G4	2M4P
T2B-1	TTGTTGGTGGGTGGGTGGGT	parallel G4	-
TB-3	TTGGGTGTGGTGGGTGGGT	parallel G4	-
G15	TTGGGGGGGGGGGGGGGT	parallel G4	2MB2
Myc1245	TTGGGGAGGGTTTTAAGGGTGGGGAAT	parallel G4	6NEB
AT11	TGGTGGTGGTTGTTGTGGTGGTGGTGGT	parallel G4	2N3M
LTR-IV	CTGGGCGGGACTGGGGAGTGGT	parallel G4	2N4Y
hras-1	TCGGGTTGCGGGCGCAGGGCACGGGCG	anti-parallel G4	-
TBA	GGTTGGTGTGGTTGG	anti-parallel G4	148D

HIV-PRO-1	TGGCCTGGGCGGGACTGGG	anti-parallel G4	-
22CTA	AGGGCTAGGGCTAGGGCTAGGG	anti-parallel G4	-
Bm-U16	TAGGTTAGGTTAGGTUAGG	anti-parallel G4	-
c-kit*	GGCGAGGAGGGGCGTGGCCGGC	anti-parallel G4	6GH0
Bom17	GGTTAGGTTAGGTTAGG	anti-parallel G4	-
G4CT	GGGGCTGGGGCTGGGGCTGGGG	anti-parallel G4	-
22GGG	GGGTTAGGGTTAGGGTTAGGGT	anti-parallel G4	2KF8
19wt	GGGGGAGGGGTACAGGGGTACAGGGG	anti-parallel G4	6FTU
LWDLN 1	GGGTTTGGGTTTTGGGAGGG	anti-parallel G4	5J05
LWDLN 2	GGGGTTGGGGTTTTGGGGAAGGGG	anti-parallel G4	2M6W
LWDLN 3	GGTTTGGTTTTGGTTGG	anti-parallel G4	5J4W
ss 3	GTCGCCGGGCCAGTCGTCCATAC	single strand	-
ss 4	GTATGGACGACTGGCCCGGCGAC	single strand	-
ss 6	GACGTGTCGAAAGAGCTCCGATTA	single strand	-
ss 7	TAATCGGAGCTCTTTGACACGTC	single strand	-
RND1	CTATACGAAAACCTTTTGTATCATT	single strand	-
RND2	AATGATACAAAAGGTTTTCGTATAG	single strand	-
RND3	TAACGTTTATAATGTAGTCTCATT	single strand	-
RND4	TAATGAGACTACATTATAAACGTTA	single strand	-
RND6	GTTGTCATTGCCCCGAATAATTCT	single strand	-
RND7	GCCTTGCGGAGGCATGCGTCATGCT	single strand	-
RND8	AGCATGACGCATGCCTCCGCAAGGC	single strand	-
dT26	TTTTTTTTTTTTTTTTTTTTTTTTTT	single strand	-
ds26	CAATCGGATCGAATTCGATCCGATTG	duplex	-
ds-lac	GAATTGTGAGCGCTCACAATTC	duplex	-
Hairpin 1	GGATTCTTGGATTTTCCAAGAATCC	duplex	-
Hairpin 2	TCGGTATTGTGTTTACAATACCGA	duplex	-
Hairpin 3	AGGACGGTGTATTTTACACCGTCCT	duplex	-
d34	GTCGCCGGGCCAGTCGTCCATAC		-
	GTATGGACGACTGGCCCGGCGAC	duplex	-
d67	GACGTGTCGAAAGAGCTCCGATTA		-
	TAATCGGAGCTCTTTGACACGTC	duplex	-
RND34	TAACGTTTATAATGTAGTCTCATT		-
	TAATGAGACTACATTATAAACGTTA	duplex	-
RND78	AGAATTATTCGGGGGCAATGACAAC		-
	GTTGTCATTGCCCCGAATAATTCT	duplex	-

Table S2. T_m of some competitors collected by UV-melting^a

Name	T _m /°C	Note	Reported conformation
F21T	58.1 ^a	Probe	-
UpsB-Q3	39.4	False negative	hybrid G4
SP-PGQ-1 ^b	46.8	Not a G4; rather a duplex	hybrid G4
KRAS-22RT	32.8	False negative	parallel G4
SP-PGQ-3	34.7	False negative	parallel G4
TBA	42.1	False negative	anti-parallel G4
Bm-U16	26.8	False negative	anti-parallel G4
LWDLN3	33.0	False negative	anti-parallel G4
Bom17	30.3	False negative	anti-parallel G4
46AG	43.3	Positive control ^c	hybrid G4
Bcl2Mid	52.0	Positive control ^c	hybrid G4
25TGA	47.3	Positive control ^c	hybrid G4
Chl1	62.7	Positive control ^c	hybrid G4
LTR-III	46.5	Positive control ^c	hybrid G4
c-kit-T12T2	46.2	Positive control ^c	parallel G4
Pu24T	74.1	Positive control ^c	parallel G4
c-kit87up	48.7	Positive control ^c	parallel G4
VEGF	66.5	Positive control ^c	parallel G4
T95-2T	81.4	Positive control ^c	parallel G4
22CTA	45.5	-	anti-parallel G4
G4CT	69.5	-	anti-parallel G4
22GGG	54.4	-	anti-parallel G4
19wt	80.9	-	anti-parallel G4
LWDLN1	45.1	-	anti-parallel G4
ds26	76.3	-	duplex
ds-lac	70.1	-	duplex
Hairpin1	71.4	-	duplex
Hairpin2	74.9	-	duplex
d34	78.8	-	duplex

^a T_m of F21T was measured by FRET-melting, T_m of reported G4 sequences and duplex were collected at 295 nm and 260 nm, respectively.

^b T_m of SP-PGQ-1 calculated by UV-melting curve at 260 nm.

^c Positive controls are G4-forming sequences which give an S value close to 0.

Table S3. Tm and S Factor average of some sequences in the training set

Name	Tm/°C^a	S Factor average^b
46AG	43.3	0.49
c-kit2-T12T12	46.2	0.40
25TAG	47.3	0.60
Bcl2Mid	52.0	0.62
22GGG	54.4	0.50
G4CT	69.5	0.35
Pu24T	74.1	0.23
19wt	80.9	0.22
T95-2T	81.4	0.26

^a Tm were collected at 295 nm.

^b S Factor averages were determined by the S of 0.2 / 0.6 / 1 / 2 μ M (1x to 10x molar excess, as compared to F21T) competitor concentrations.

Table S4. Testing set of DNA sequences.

Name	Sequence (5'-3')	G4-Hunter score	Location (Human hg19)
F21T	FAM-GGGTTAGGGTTAGGGTTAGGG-TAMRA	-	-
UN1	CGGGCAGGGAGGGCGGCTGTGCGGG GC	1.59	chr3: 196045150-196045176
UN2	TGGGGCGGGGAAGAGGGGCGGGG T	2.73	chr8: 57124113-57124138
UN3	CGGGAAGGGGCGGGCGCAATGGGC	1.62	chr17: 62915412-62915435
UN4	TGGGAGGCGGAGGTGGGCAGGTTGCT	1.23	chr17: 16258771-16258791
UN5	GTGCTGGGGCGCCCACTTCGGGGTGG TGC	0.90	chr11: 364605-364633
UN6	AGTTGGTAGGCTGAGGCGGGAGGATT GC	0.89	chr5: 64905453-64905472
UN7	AGGGCCGGGAGAGGGATCCGCCATAT TGGAGCTGGGGC	0.94	chr14: 36278253-36278290
UN8	AGGAAGCTGGGGTAGGAGAATTGCTTG A	0.93	chr12: 57876780-57876802

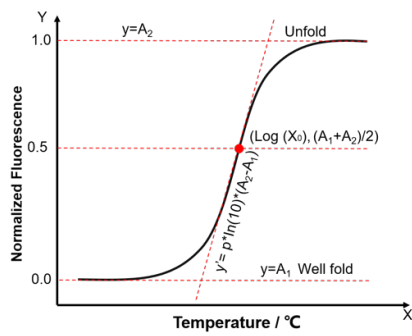
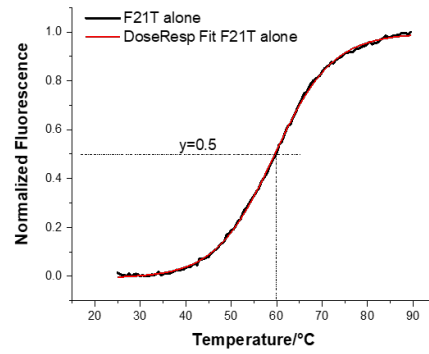
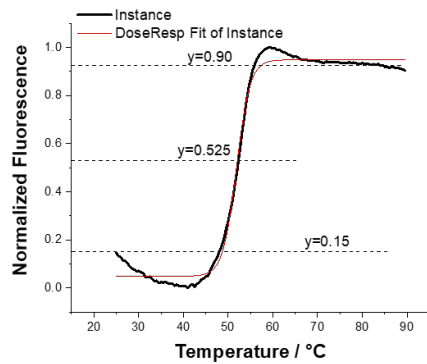
A**B****C**

Fig. S1. Different methods to calculate T_m . (A) Sample curve of DoseResp. (B) Two methods to calculate T_m of F21T alone. (C) Example of an atypical FRET-melting curve, for which accurate T_m determination is more difficult.

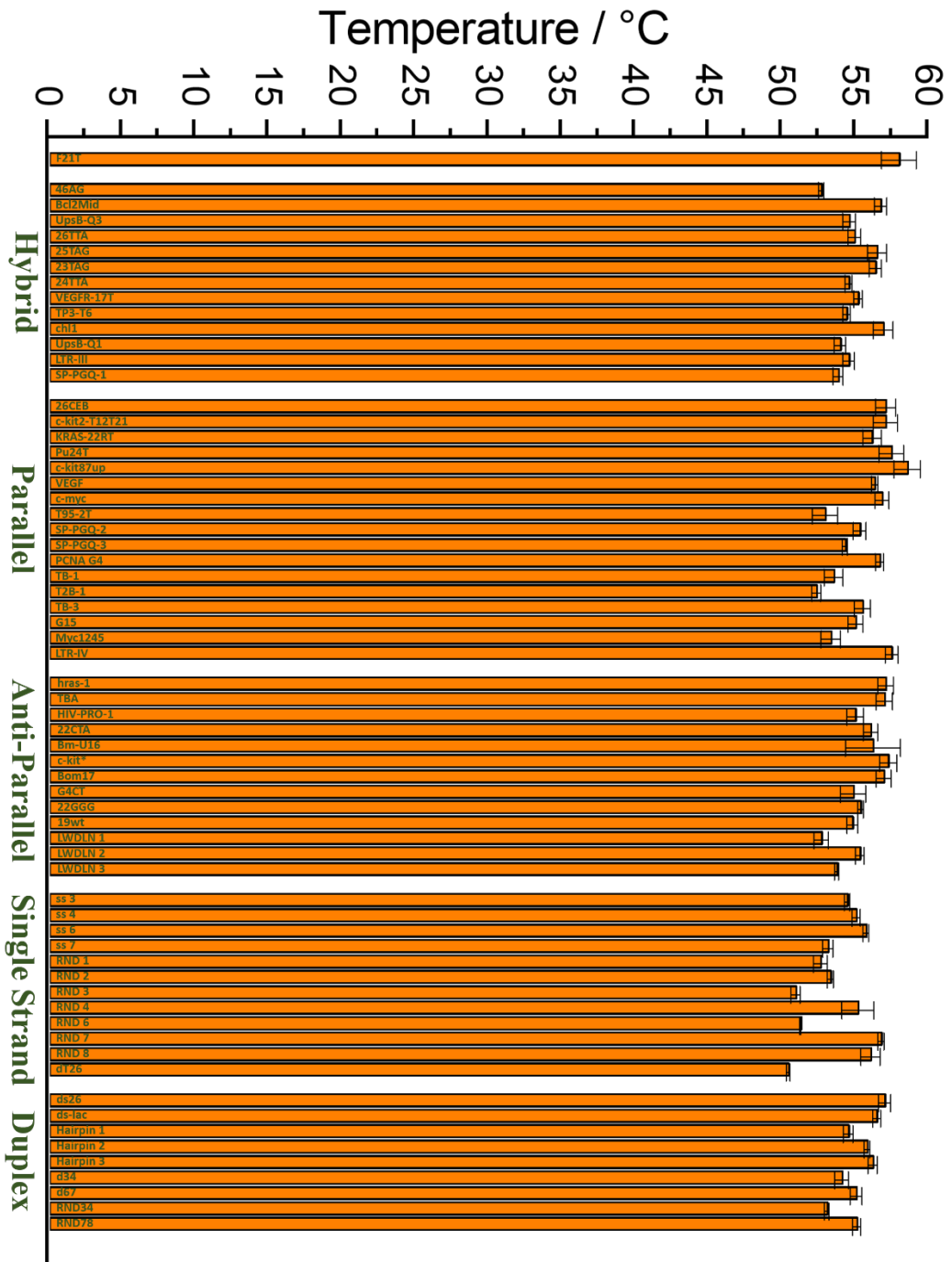


Fig. S2. FRET Tm of 0.2 μ M F21T alone (top) or in the presence of various competitors.

All competitors were tested at 3 μ M strand concentration.

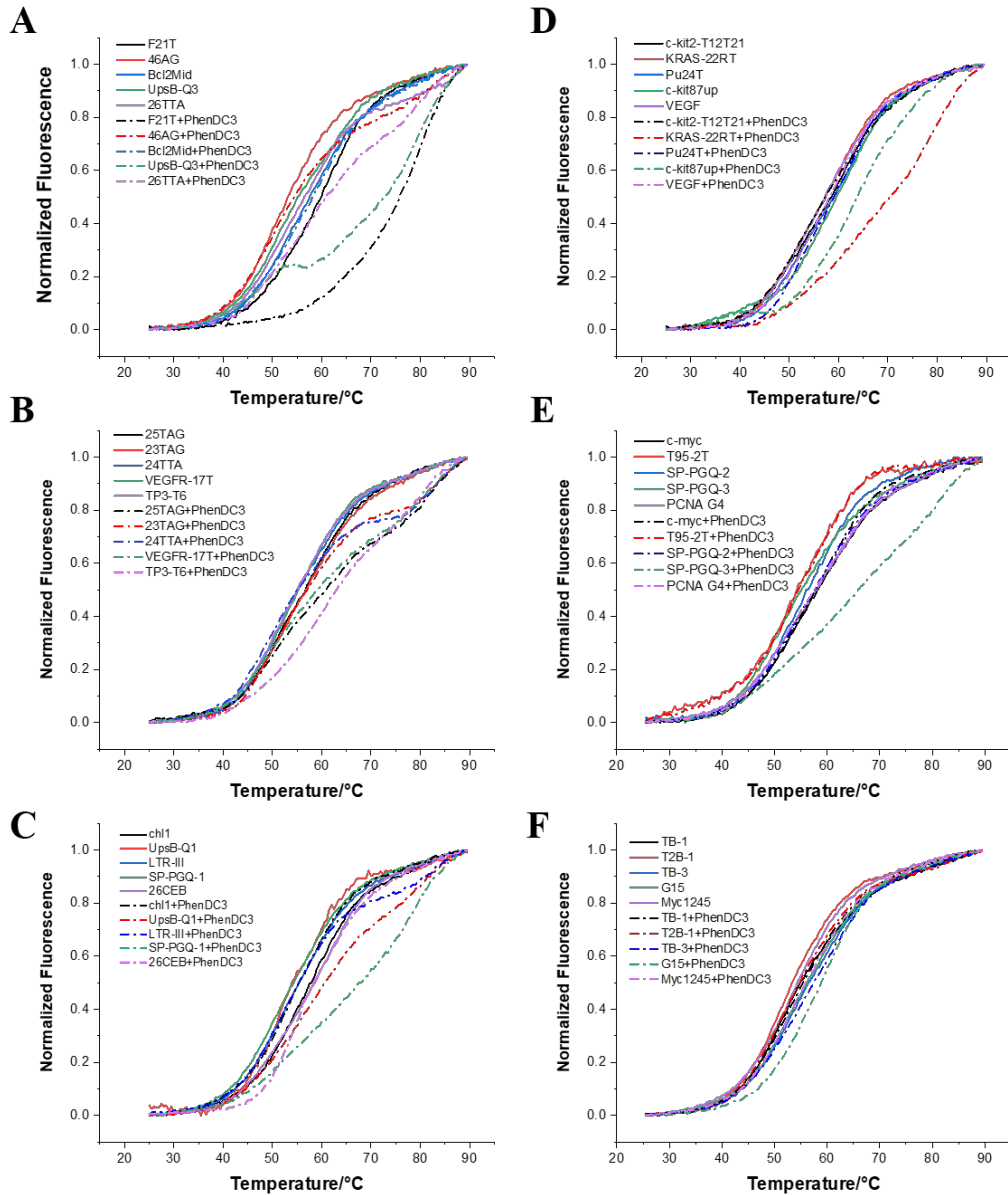


Fig. S3. Normalized FRET-melting curves of 0.2 μ M F21T in the presence of various competitors (3 μ M strand concentration) used in the training set, with or without 0.4 μ M PhnDC3. Samples were annealed and measured in 10 mM KCl, 90 mM LiCl, 10 mM lithium cacodylate pH 7.2 buffer.

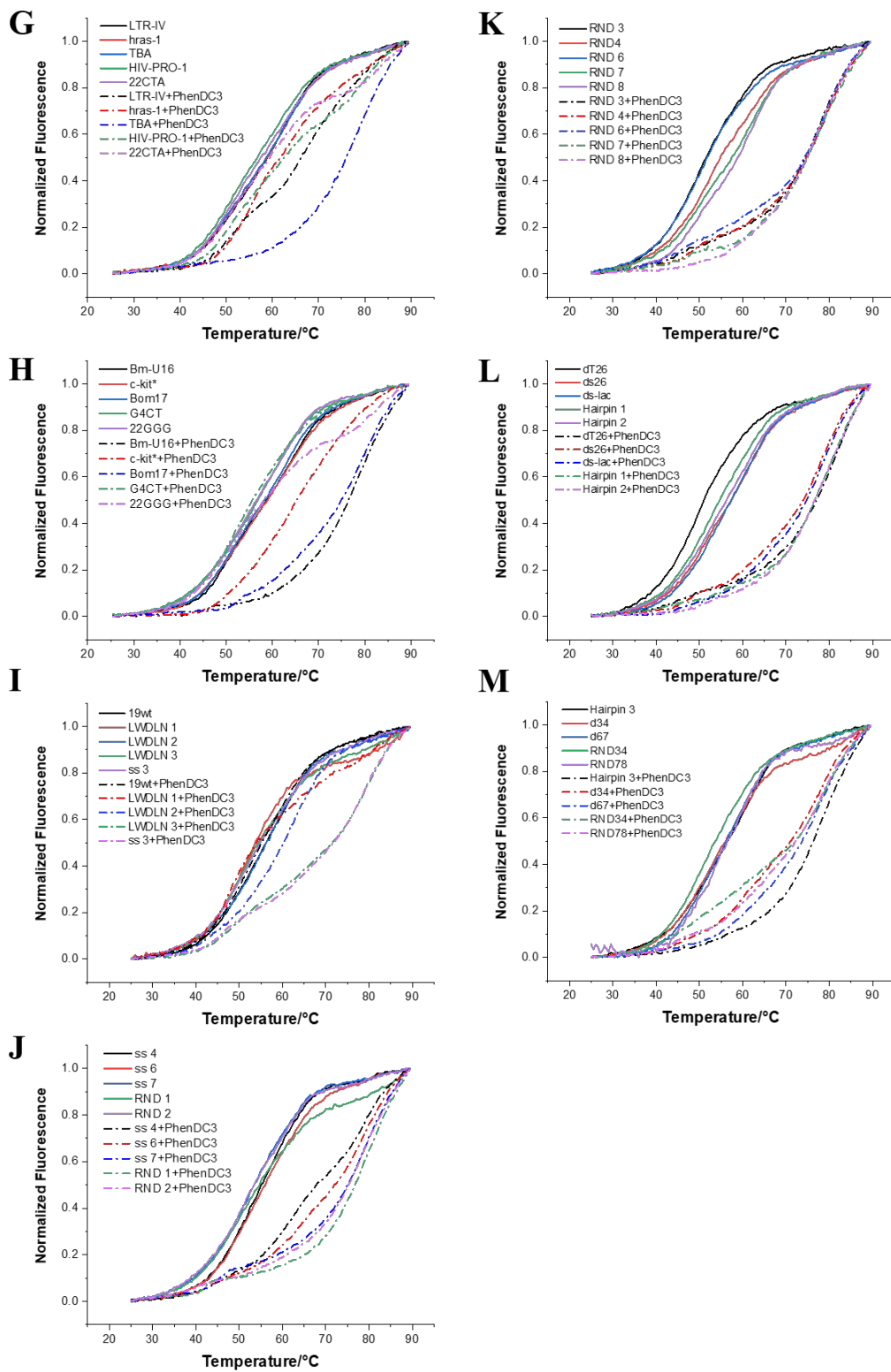


Fig. S3, continued.

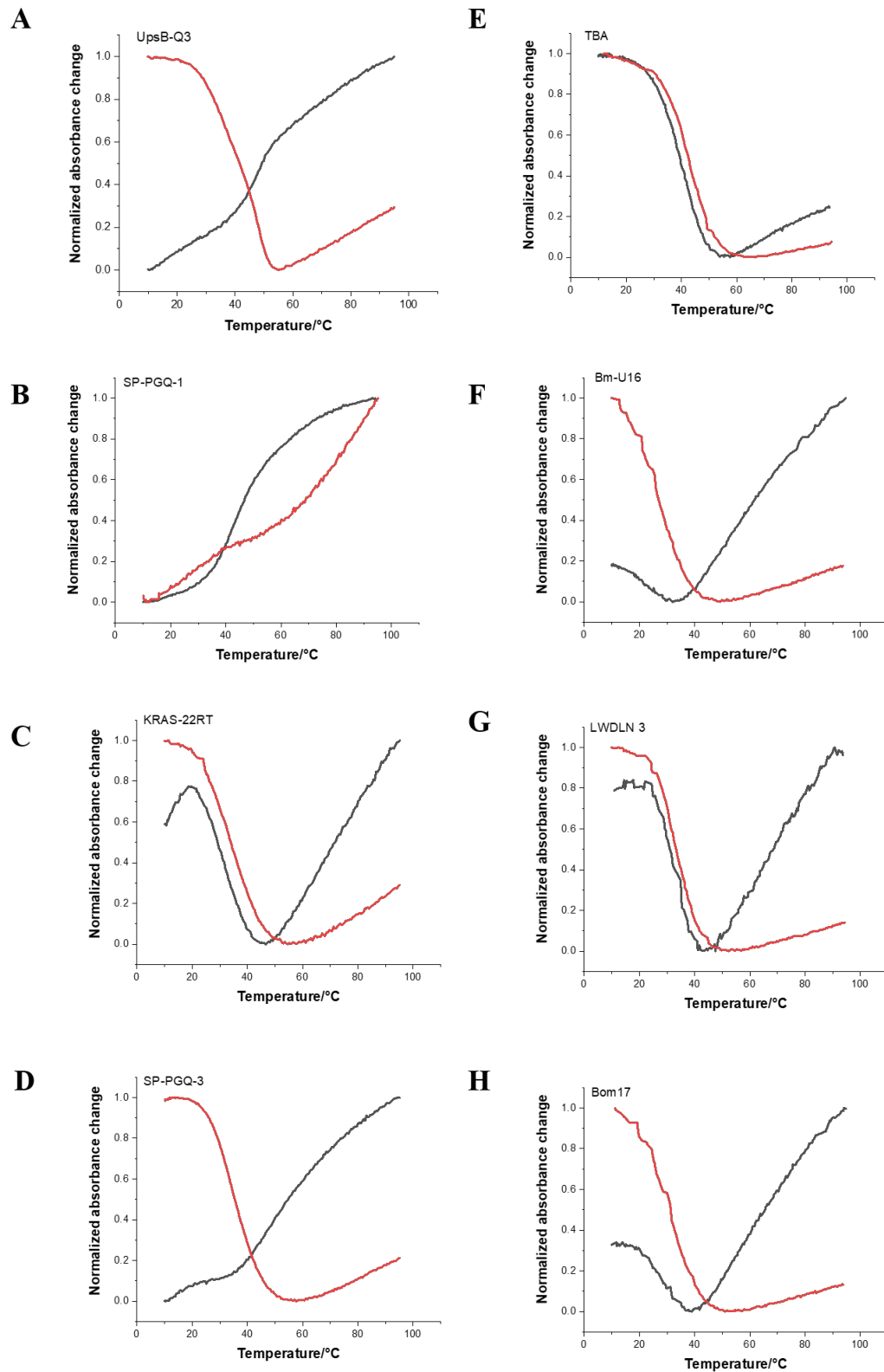


Fig. S4. Normalized UV-melting curves for some sequences of the training set. All data collected in 10-95 °C, monitored at 260 nm (Blank lines) and 295 nm (Red lines).

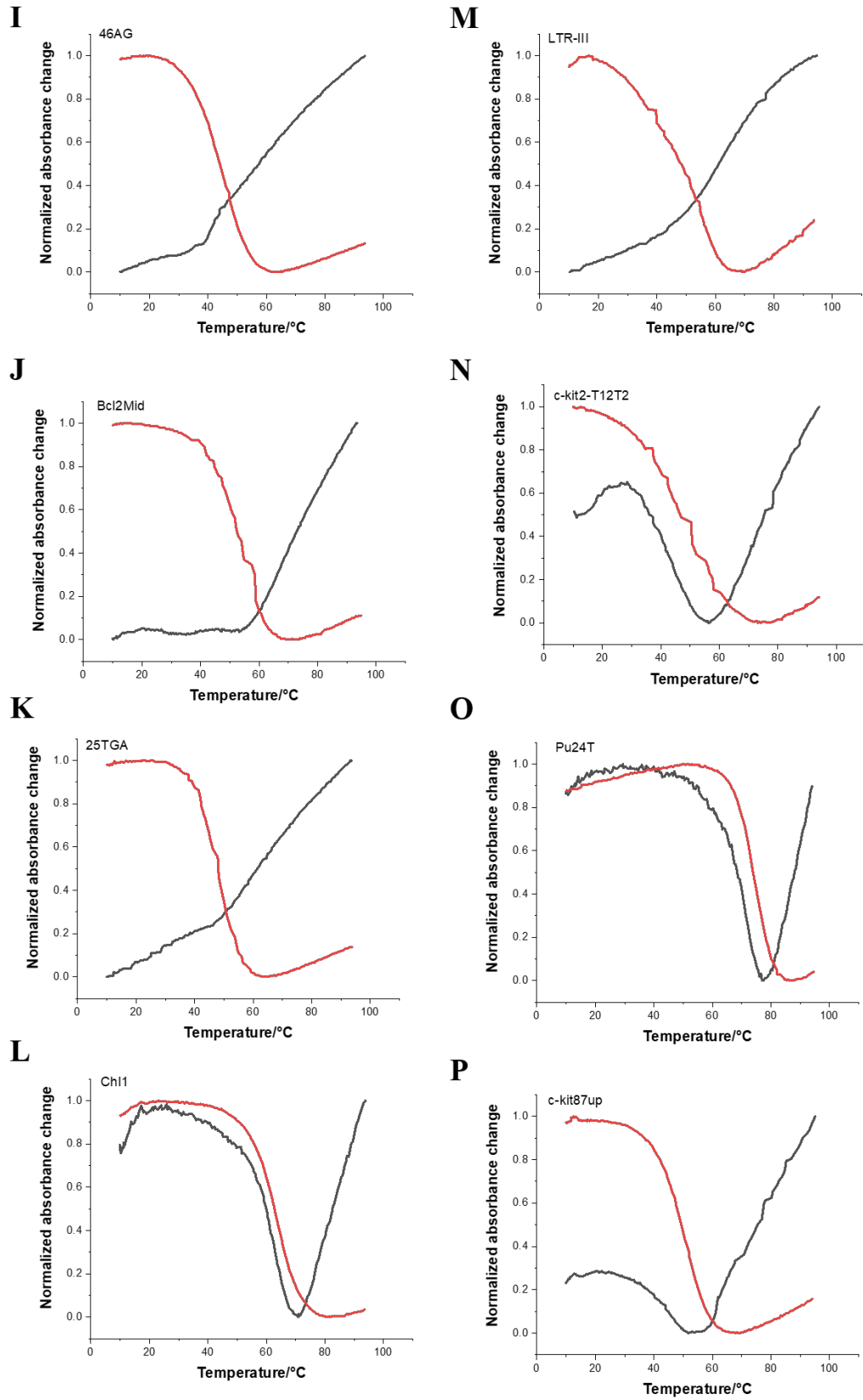


Fig. S4. *continued.*

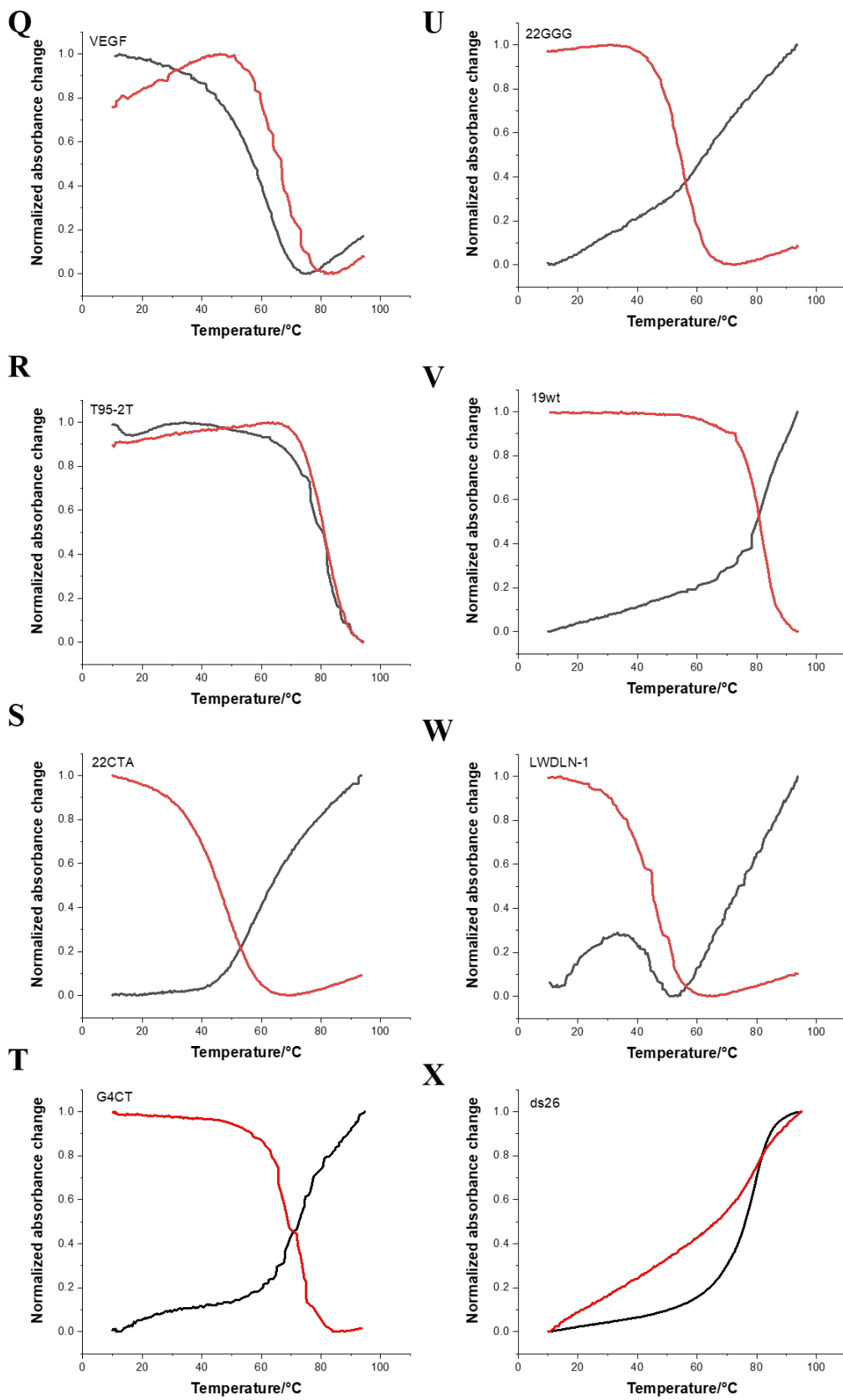


Fig. S4. *continued.*

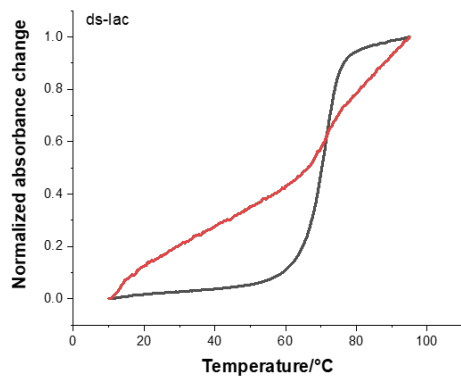
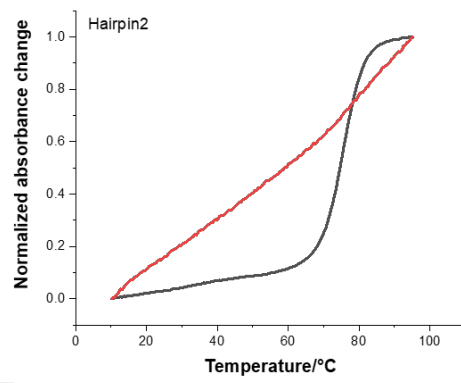
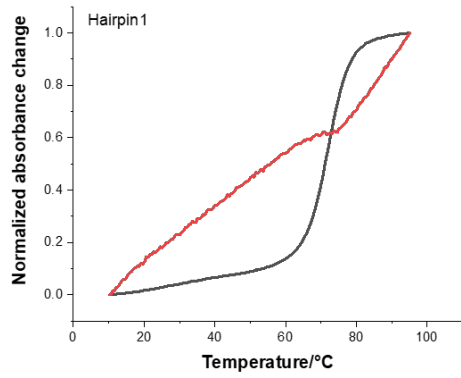
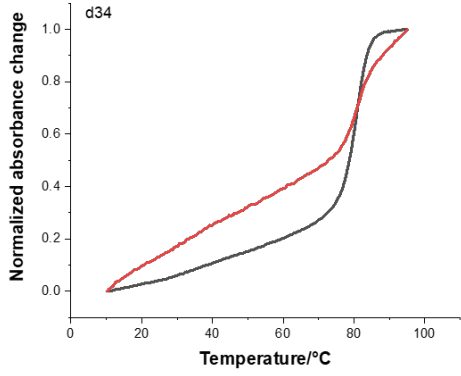
Y**AA****Z****AB**

Fig. S4. *continued.*

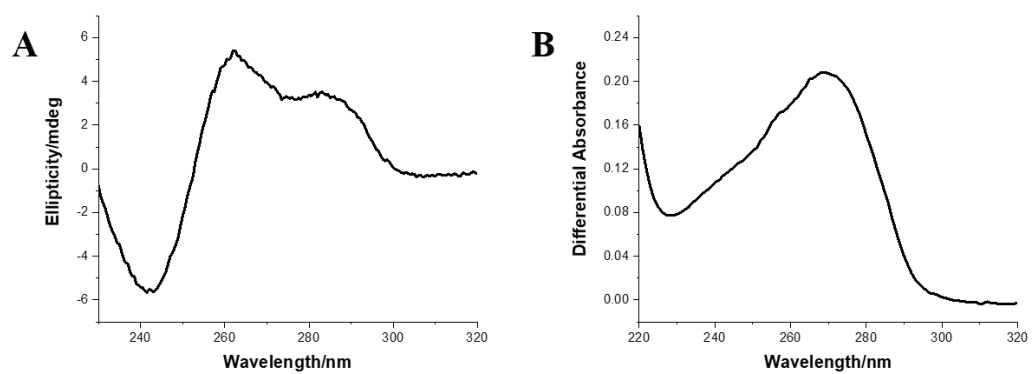


Fig. S5. CD **(A)** and thermal differential absorbance (TDS) **(B)** spectra for 3 μ M SP-PGQ-1 in the FRET buffer.

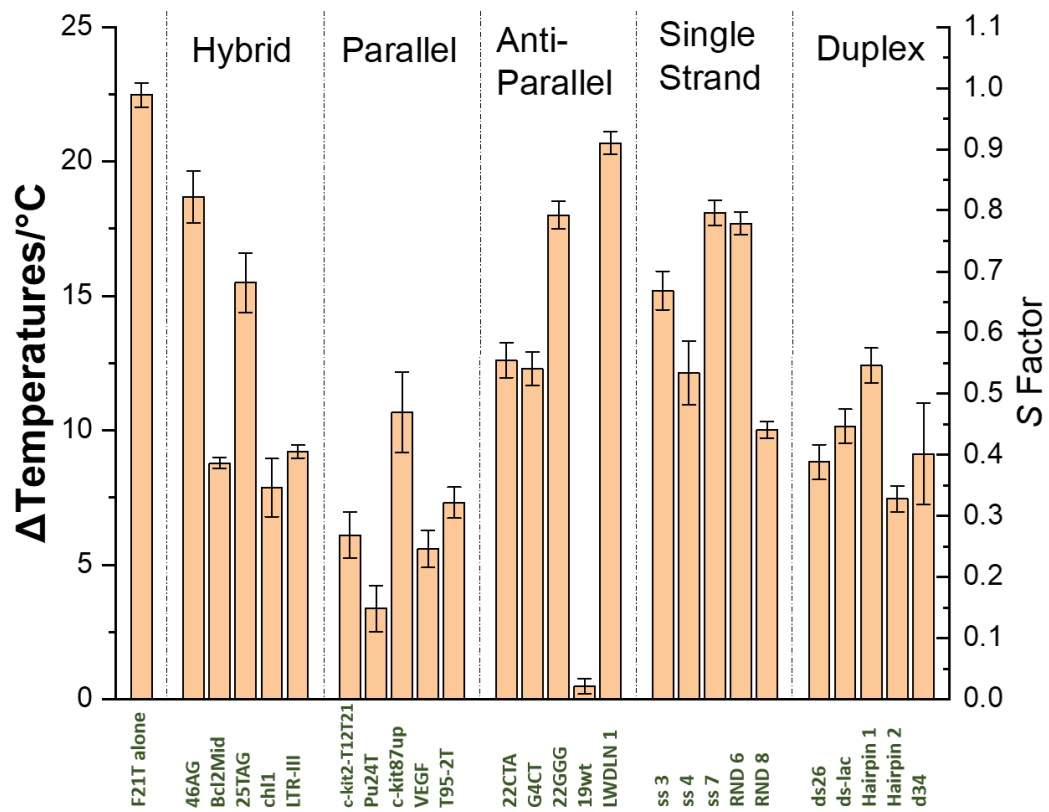


Fig. S6. ΔT_m induced by 0.4 μM TMPyP4 on 0.2 μM F21T, alone or in the presence of 3 μM competitors. The *S Factor* is also provided on the right Y-axis. Samples were annealed and measured in 10 mM KCl, 90 mM LiCl, 10 mM lithium cacodylate pH 7.2 buffer.

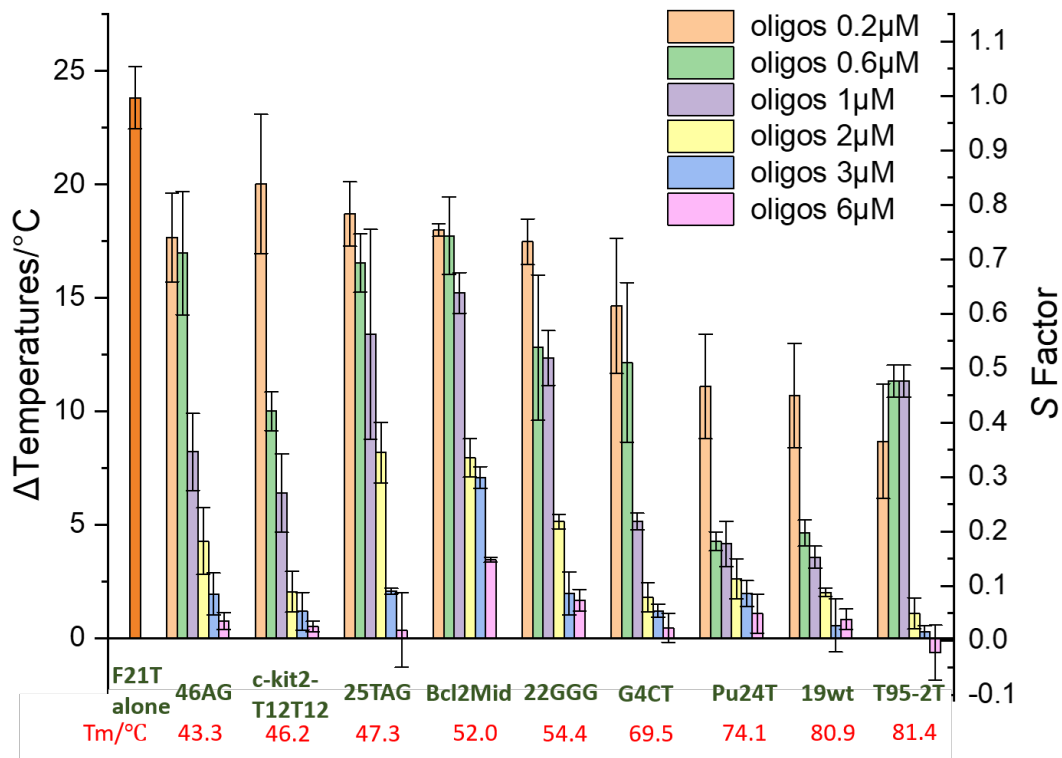
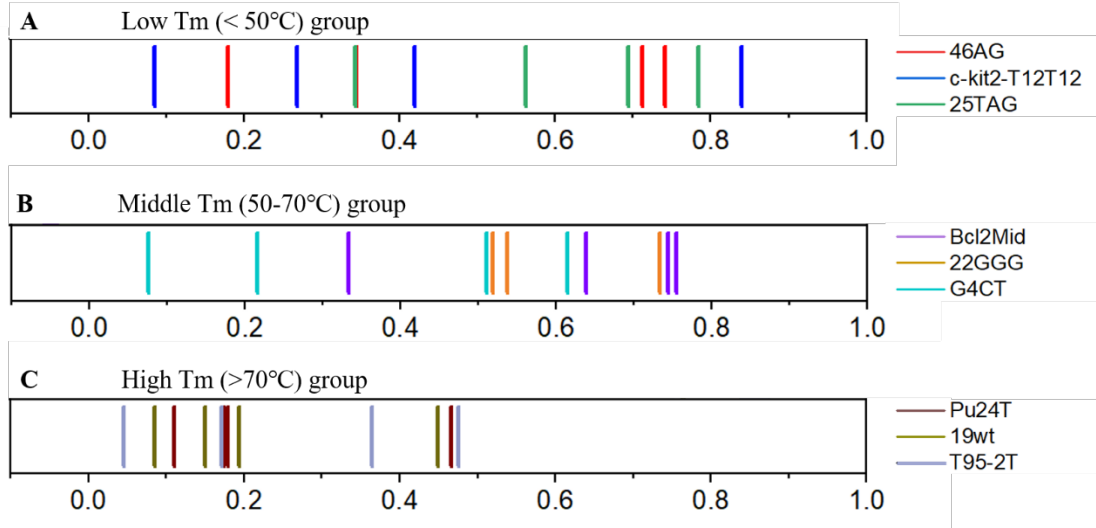


Fig. S7. ΔT_m induced by 0.4 μM PhenDC3 on 0.2 μM F21T, alone or in the presence of 0.2 / 0.6 / 1 / 2 / 3 / 6 μM competitors (1x to 30x molar excess, as compared to F21T). The *S Factor* is also provided on the right Y-axis. Samples were annealed and measured in 10 mM KCl, 90 mM LiCl, 10 mM lithium cacodylate pH 7.2 buffer. The T_m for each quadruplex are shown in red below; sequences were ranked from left (lowest T_m) to right (highest T_m).



Distribution of S Factor

Fig. S8. Distribution of S-factor values for 0.2 / 0.6 / 1 / 2 μM competitor concentrations (1x to 10x molar excess, as compared to F21T). Oligonucleotides are grouped based on T_m : low T_m ($< 50^\circ\text{C}$) (**A**), middle T_m ($50\text{-}70^\circ\text{C}$) (**B**) and high T_m ($>70^\circ\text{C}$) (**C**). The 4 S values shown for each competitor are depicted with the same color; the lowest S value (leftmost vertical bar) corresponds to the highest (2 μM) concentration while the highest S value (rightmost bar) corresponds to the lowest (0.2 μM) concentration

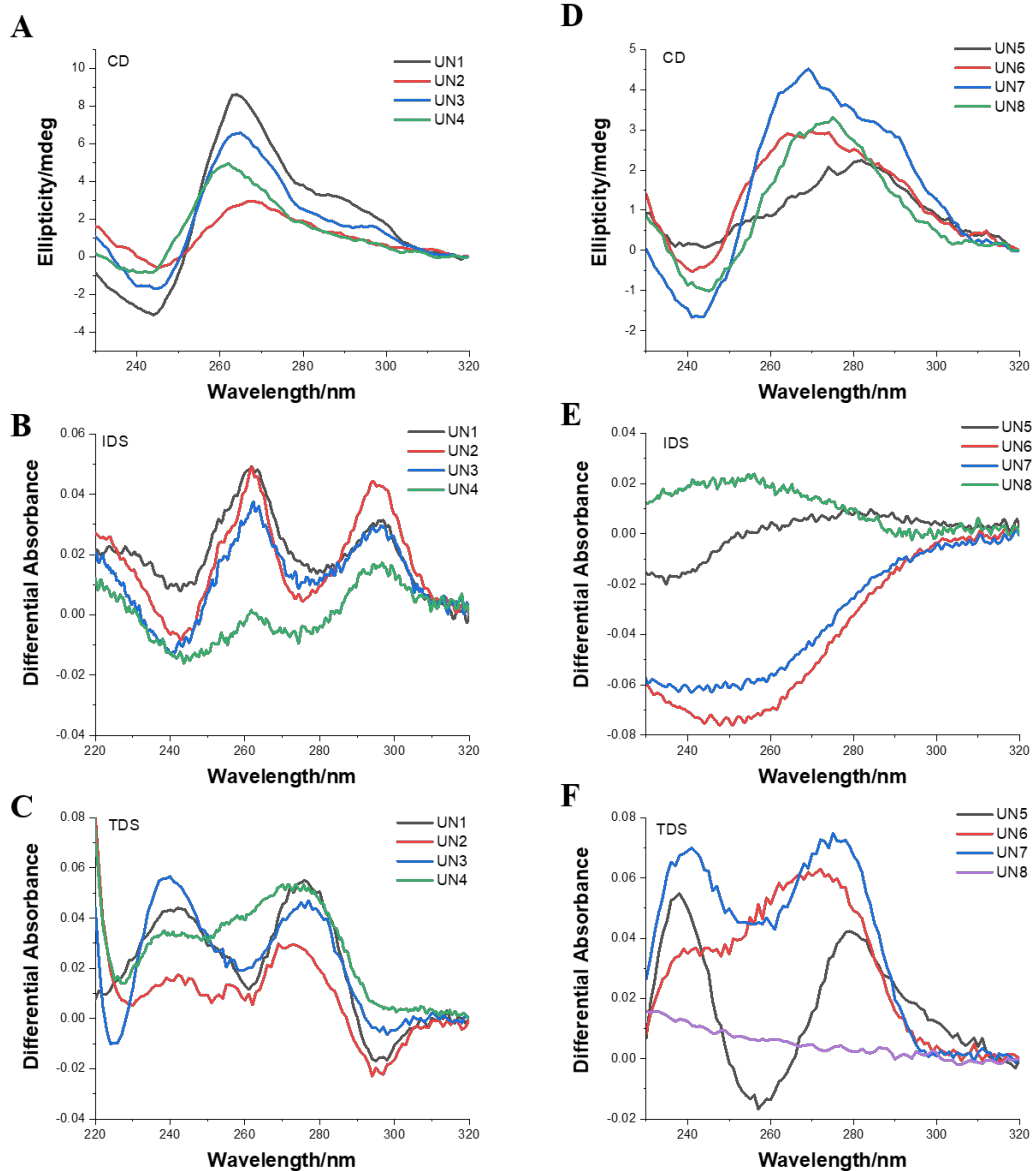


Fig. S9. CD, IDS, and TDS spectra of testing set. Panels **A, B, C** correspond to positive sequences (absolute S values < 0.1) and panels **D, E, F** to negative sequences, unable to compete ($S > 0.8$). From top to bottom: CD spectra, IDS spectra, and TDS spectra.

Chapter III. Isothermal FRET competition assay

F22 and 37Q are natural partially complementary RNA strands, originated from the human telomerase RNA (*hTERC*). The F22 - 37Q fluorescent system was first employed to characterize if a compound had affinity to G4 structures. As shown in **Figure 15**, 37Q is a G-rich strand able to fold into a G4 that can be recognized by G4 ligands, preventing hybridization to F22 and letting on the fluorescence of the latter. Differently, when 37Q hybridizes to F22, FAM fluorescence is quenched. In other words, the duplex-quadruplex equilibrium of 37Q can be shifted by G4 ligands, and the fluorescence intensity of F22 depends on the affinity of the compound and 37Q G4 formation. The FRET-based assay can be performed in microwell plates, which are sample- and time-saving.

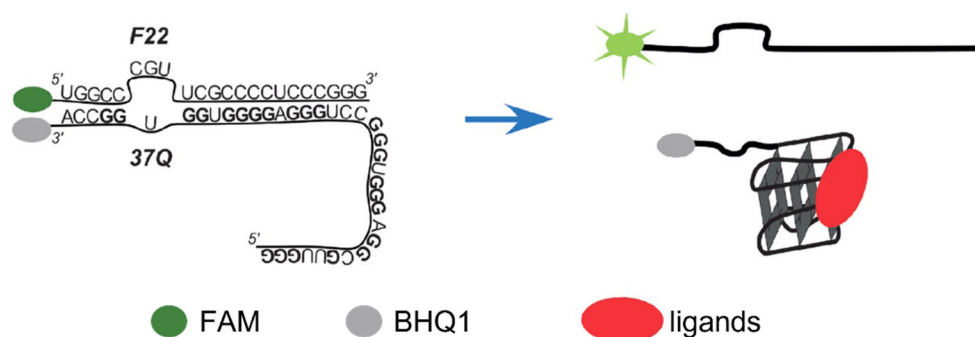


Figure 15 Principle of the original assay for G4 ligand characterization. F22 hybridizes with 37Q in the absence of G4 ligands, resulting in fluorescence quench; whereas the presence of G4 ligands inhibits F22-37Q duplex formation, thus the fluorescence of F22 can be observed. The figure was taken from [44].

Although FRET-MC is an inexpensive and fast G4 characterization method, the heating process, on which the method is based on, leads to misestimation of G4-forming competitors with poor thermal stability. The duplex-G4 equilibrium provides a new approach to design an isothermal FRET assay, and overcome the disadvantage of FRET-MC. Similar to FRET-MC, the novel iso-FRET is also based on the competitive binding of PhenDC3 between the G4-forming quencher (37Q) and an unknown competitor. Fluorescent F22 is then added to reveal the conformation of 37Q: if the competitor adopts a G4 structure and traps PhenDC3, 37Q forms a duplex with F22 and quenches F22 fluorescence; differently, if PhenDC3 stabilizes 37Q, the latter can no longer hybridize to F22, letting on the fluorescence of F22. Both competitive interaction and hybridization occur spontaneously at room temperature (25 °C), and our results also showed that the iso-FRET assay works well at human physiological temperature (37 °C). By using the iso-FRET assay, it is theoretically possible to detect G4 competitors characterized by affinities as lower as 17 μ M toward PhenDC3, which is much lower than the equilibrium dissociation constant (K_d) of PhenDC3 towards G4s (nM range) reported in literature [251, 252]. The risk of a false negative in the iso-FRET assay is therefore

extremely low.

The ionic strength of the iso-FRET buffer affects F22 + 37Q hybridization, the folding of G4 competitors, and the competition between 37Q and competitors for PhenDC3 binding. To choose the most appropriate buffer for the iso-FRET assay, we kept ionic strength constant at 100 mM, and adjusting the ratio of potassium and lithium. The difference between G4 and non-G4 competitor was maximized in the 20 mM KCl, 80 mM LiCl, 10 mM LiCaco buffer (pH =7.2). '20K' buffer was therefore chosen for all subsequent experiments.

Iso-FRET assay is a kinetic-controlled method. We found that the hybridization of F22 and 37Q is much slower (hours long) than regular duplexes; the relatively long equilibration time is caused by 37Q quadruplex formation. Hence, the incubation time after the addition of F22 is an important element of the iso-FRET assay, especially in the presence of some weak G4 competitors. SP-PGQ-3, TBA, Bm-U16, Bom 17 and LWDLN3 exhibited a time-dependent behavior: they showed relatively high fluorescence at incubation times shorter than 3 h (**Figure 16a**). The weak G4s can be correctly characterized by extending the incubation time to 24 h (**Figure 16b**). Fluorescence intensities were normalized with the *F value*, which reflected the amount of PhenDC3 bound to 37Q. Boundaries between G4 and non-G4 competitors with 95% prediction interval were calculated: *i*) $F < 0.33$: G4 competitor; *ii*) $0.33 \leq F < 0.54$: unknown; *iii*) $F \geq 0.54$: non-G4 competitor. Of note, the boundaries may fluctuate between different batches of experiments. Any iso-FRET result obtained near the boundaries should be treated with caution.

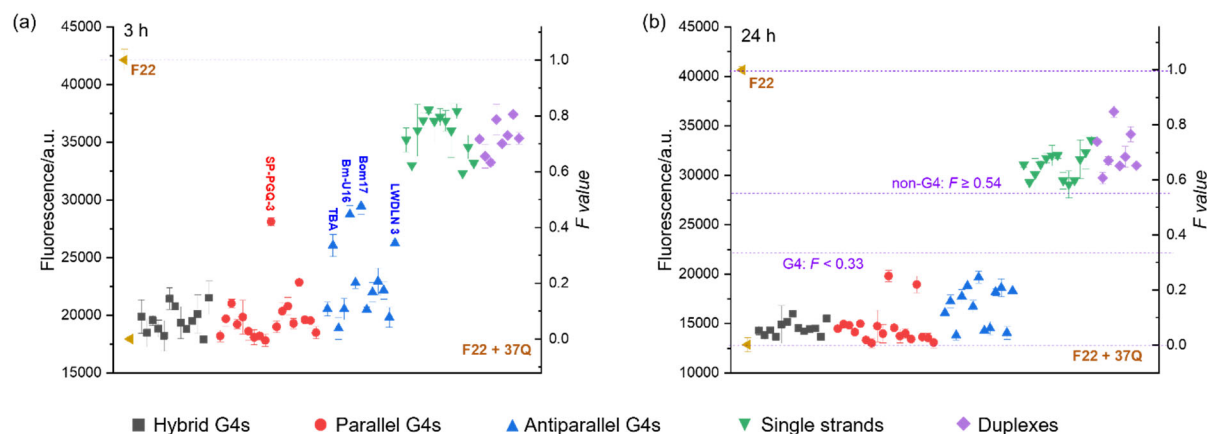


Figure 16 Fluorescence quenching in the presence of various competitors. F22 is incubated for (a) 3 h or (b) 24 h in the presence of 37Q and PhenDC3 alone or in the presence of competitors. The *F values* (right Y-axis) provide a normalized value. The four different horizontal dotted lines in panel (b) from bottom to top correspond to (i) the level of fluorescence in the absence of a competitor ($F = 0$); (ii) the first threshold value at 0.33 chosen for positive samples (G4-forming sequences all exhibit *F values* between 0 and 0.33); (iii) the second threshold value at 0.54: negative controls/non G4-forming sequences all exhibit *F values* between 0.54

and 1; (iv) the level of fluorescence of F22 alone, with no 37Q added ($F = 1$). Samples were measured in 20 mM KCl, 80 mM LiCl, 10 mM LiCaco buffer (pH = 7.2) at 25 °C.

In addition to addressing the false negative problem identified in the FRET-MC assay for G4 competitors with poor thermal stability, iso-FRET simplified data handling and increased the throughput from 96 to 384 samples per plate [250]. However, iso-FRET is still not a perfect method for G4 characterization: it cannot be used to test competitors which have a high complementarity to the C-rich probe strand F22. In the presence of a large excess of a competitor with high complementarity towards F22, a competitor-F22 duplex may form and prevent fluorescence quenching, leading to false negative results. In FRET-MC, the complex formed by the competitor and F21T can be easily detected by the generation of a reverse melting curve, while iso-FRET only gives the final fluorescence intensity. Iso-FRET does not give any information about competitor-F22 complex formation. Therefore, we defined the *CF factor* based on the global base pairing alignment, to describe the complementarity degree between the competitor and F22. $CF = 0$ means the competitor is absolutely devoid of base pairing to F22, while a competitor perfectly complementary to F22 gives the highest *CF* value of 1. *CF factor* is easy to calculate, however, it is not particularly precise: A-T and C-G bps are considered equally important, but in fact, A-T bps contribute less to the duplex stability than C-G bps, as the A-T bp contains two H-bonds while C-G bp involves three H-bonds. Our main purpose is not to study the weight of A-T and C-G bps during the hybridization process; until now, the *CF* factor determined solely according to the simple alignment has not met any issue. We evidenced that the iso-FRET assay is able to identify DNA competitor structures with *CF factor* lower than 0.86. We strongly recommend to check *CF factor* of competitors before testing, to avoid using the iso-FRET to competitors with extremely high *CF*.

Iso-FRET: an isothermal competition assay to analyze quadruplex formation *in vitro*

Yu Luo^{1,2}, Daniela Verga^{2,3} and Jean-Louis Mergny^{1,*}

¹Laboratoire d'Optique et Biosciences, Ecole Polytechnique, CNRS, Inserm, Institut Polytechnique de Paris, 91128 Palaiseau, France, ²CNRS UMR9187, INSERM U1196, Université Paris-Saclay, F-91405 Orsay, France and ³CNRS UMR9187, INSERM U1196, Institut Curie, PSL Research University, F-91405 Orsay, France

Received October 26, 2021; Revised April 26, 2022; Editorial Decision May 13, 2022; Accepted May 16, 2022

ABSTRACT

Algorithms have been widely used to predict G-quadruplexes (G4s)-prone sequences. However, an experimental validation of these predictions is generally required. We previously reported a high-throughput technique to evidence G4 formation *in vitro* called FRET-MC. This method, while convenient and reproducible, has one known weakness: its inability to pin point G4 motifs of low thermal stability. As such quadruplexes may still be biologically relevant if formed at physiological temperature, we wanted to develop an independent assay to overcome this limitation. To this aim, we introduced an isothermal version of the competition assay, called iso-FRET, based on a duplex-quadruplex competition and a well-characterized bis-quinolinium G4 ligand, PhenDC3. G4-forming competitors act as decoys for PhenDC3, lowering its ability to stabilize the G4-forming motif reporter oligonucleotide conjugated to a fluorescence quencher (37Q). The decrease in available G4 ligand concentration restores the ability of 37Q to hybridize to its FAM-labeled short complementary C-rich strand (F22), leading to a decrease in fluorescence signal. In contrast, when no G4-forming competitor is present, PhenDC3 remains available to stabilize the 37Q quadruplex, preventing the formation of the F22 + 37Q complex. Iso-FRET was first applied to a reference panel of 70 sequences, and then used to investigate 23 different viral sequences.

INTRODUCTION

Different from the classical double-helix, G-quadruplexes (G4) constitute a family of specific DNA and RNA secondary structures. G4s result from the stacking of two or more G-quartets, *i.e.* planar layers of four guanines held together by Hoogsteen hydrogen bonding (1,2). G-quadruplexes have been studied from different perspectives

and have been found in a variety of genomes, including pathogens (*e.g.* Nipah (3) and Ebola (4) viruses), bacteria and Archaea (5), as well as many eukaryotes, where G4 motifs are often found in promoters (6) and close to the origins of replication for mammals (7). G-quadruplexes play important roles in biological processes, including genome stability (8,9), regulation of gene expression (10–12), specific chromatin remodeling and replication (13,14), and RNA metabolism (1). How to find potential G4 motifs and then characterize their structures are basic questions concerning G-rich sequences in genomes. In 2016, we introduced a novel prediction method, G4Hunter, to discover potential G-rich sequences located within genes or genomes, also able to provide a rough estimation about the possibility of the target sequence to form a G4 (15). G4Hunter has been used to find G-rich sequences in a variety of species, including human (16), *Plasmodium* (17), *Dictyostelium* (18) and viral (19,20) genomes, and is now available as a web application (21).

Compared to model G4 structures, natural G4 sequences are often more complex and irregular. Genomic G4 motifs vary in length and may form a number (or variety) of non-canonical topologies. The first bioinformatic approaches performed in 2005 estimated that there were >300 000 G4-prone sequences in the human genome (22,23). With the development of sequencing technologies, 736 689 G4 structures have been identified *in vitro* (24). Even if the real number of G4-motifs actually formed may be lower than the one so far identified, most biophysical approaches are unable to deal with so many candidate sequences. For example, high resolution structural methods such as nuclear magnetic resonance (NMR) (25) cannot easily handle hundreds or thousands of samples. For this reason, rapid high-throughput assays able to deal with hundreds of motifs are needed.

Previously, we developed a thermal competition assay, the so-called FRET-MC assay, to characterize if an unknown sequence forms a quadruplex (26). The two ends of a short single-stranded DNA oligonucleotide mimicking ≈4 copies of the human telomeric motif, Tel21, were labeled with fluorescein (FAM) and TAMRA. This FAM-Tel21-TAMRA sequence, hereafter abbreviated to F21T, was used

*To whom correspondence should be addressed. Tel: +33 169335001; Email: jean-louis.mergny@inserm.fr

as a fluorescent probe sequence in the FRET-MC assay, which is based on the competitive binding of a selective G4 ligand between F21T and an unknown competitor. In principle, any G4-forming sequence would act as decoy for a specific G4 ligand, meaning that less compound would be available to stabilize F21T. This assay is highly reliable, with one weakness: it fails to identify G4-motifs with a low thermal stability as they would be single-stranded at the temperature where F21T unfolds.

For this reason, we wished to develop an isothermal version of this competition assay, which would overcome the issues caused by differences in thermal stabilities. To this aim, we designed a duplex-quadruplex competition assay derived from the system developed by Lacroix *et al.* (27), with a pair of probe strands consisting of a quencher (37Q, a strand from a telomeric sequence *hTERC*) and a partially complementary strand labeled with FAM and named F22 (27). In a manner similar to FRET-MC, a well-characterized and highly specific G4 ligand such as PhenDC3 (28) was used. The sequence of interest (competitor X) would compete with 37Q for PhenDC3 binding if, and only if, it adopts a quadruplex structure. The difference with FRET-MC is that the isothermal system here relies on a duplex-quadruplex competition: hybridization between 37Q and F22 would occur only if PhenDC3 is not present, or unavailable due to binding with the sequence to be tested. All steps are processed at a constant temperature, allowing true high-throughput. We validated this iso-FRET assay with a training set of DNA and RNA sequences containing positive (quadruplexes of different stabilities and topologies) and negative (single-strands and duplexes) controls. Kinetic considerations, advantages and disadvantages of this method are discussed.

MATERIALS AND METHODS

Samples

Non-labeled oligonucleotides were purchased RP cartridge purified from Eurogentec (Seraing, Belgium) as dried samples. Fluorescently labeled oligonucleotides (F22, F22m, 37Q, 37Qm and Cy5-37merR) were purchased from IBA (Göttingen, Germany). All sequences are provided in Supplementary Tables S1–S3. Except for salmon sperm DNA for which concentration is expressed in nucleotides, all other DNA/RNA concentrations were expressed as strand concentrations. DNA samples were stored at -20°C and RNA samples were kept at -80°C . Stock solution of PhenDC3 was prepared in DMSO at 2 mM concentration and stored at -20°C .

Determination of the equilibrium constant (K_d) between a G-quadruplex and PhenDC3

The K_d was measured as described by Le *et al.* (29). The K_d measurement was performed in 96-well plates with a Tecan Infinite M1000 Pro plate reader (France). A cyanine fluorophore (Cy5) was attached to the 5' end and the fluorescently labeled strand was named as Cy5-37merR. 5 μl of 100 nM Cy5-37merR and 5 μl of PhenDC3 were added to give final concentrations in each well of 10 nM for Cy5-37merR and 0/2.5/5/7.5/10/25/50/75/100/250/500/750/1,000/2,500/5,

000/7,500/10,000 nM for PhenDC3, in 20K buffer (20 mM KCl, 80 mM LiCl, 10 mM lithium cacodylate, pH 7.2) with 0.4% (v/v) DMSO. The final volume was 50 μl . Plates were kept at room temperature (RT, around 25°C) for 2 h before measurements. Cy5 was excited at 633 nm and the emission wavelength was set at 647 nm, excitation and emission bandwidths were set to 5 nm, with an integration time of 20 μs . Each experimental condition was tested at least in triplicate. Fluorescence quenching was used to normalize the measurements in terms of % bound. The K_d was processed with a single-site binding model (GraphPad Prism V 8.4.2) for curve fitting.

Kinetics

Kinetics experiments were performed in 96-well plates with a Tecan Infinite M1000 Pro plate reader (France).

For 37Q folding: (i) 200 nM 37Q was kept in 25 μl of 10 mM lithium cacodylate buffer (pH 7.2) containing potassium at different concentrations; or (ii) 200 nM 37Q in 25 μl of corresponding buffers contained 1 μM salmon sperm DNA and 1 μM PhenDC3 (0.4% v/v DMSO). Absorbance was recorded at 295 nm, interval time was set at 7.5 s, with 200 kinetics cycles and a settle time of 0 ms.

For hybridization: 250 nM 37Q (or 37Qm) was kept in 20 μl of 10 mM lithium cacodylate buffer (pH 7.2) containing potassium at different concentrations for 5 min, then 5 μl of 100 nM F22 (or F22m) were added to the corresponding buffer. The fluorophore (FAM) attached to F22 (or F22m) was excited at 492 nm and the emission wavelength was set at 520 nm, excitation and emission bandwidths were set to 10 nm, with an integration time of 20 μs . Each experimental condition was tested at least in triplicate.

Isothermal FRET competition assay

Iso-FRET was performed in 96-well plates with a Tecan Infinite M1000 Pro plate reader (France). For competitor samples: To 5 μl of 25 μM competitor oligonucleotide mixed with 5 μl of 1 μM 37Q in corresponding buffer, 10 μl of 2.5 μM PhenDC3 were added and left to stand for 5 min. Then 5 μl of 100 nM F22 were added. Final concentrations in each well were: 5 μM competitor, 200 nM 37Q, 1 μM PhenDC3 and 20 nM F22 containing 0.4% (v/v) DMSO. Control samples contained 20 nM F22 in the presence or absence of 200 nM 37Q in 25 μl of 20K buffer. Plates were kept at RT or in an incubator at 37°C before measurements. Fluorescence measurement settings were the same as described the above.

Data analysis

F value. We first defined the *F value* parameter to evaluate the extent to which a competitor affects F22 + 37Q hybridization in the presence of PhenDC3, which is related to the fluorescence intensities (FI) of F22 alone [FI F22], F22 in the presence of 37Q [FI (F22 + 37Q duplex)] and F22 in the presence of 37Q, PhenDC3 and X, abbreviated as [FI competitors]:

$$F \text{ value} = \frac{\text{FI competitors} - \text{FI (F22 + 37Q duplex)}}{\text{FI F22} - \text{FI (F22 + 37Q duplex)}}$$

We then defined a threshold to distinguish between G4s and non-G4s competitors. *F* values were separated into the G4 group and the non-G4 group depending on the competitor structure. χ^2 test was employed to check if these two groups displayed a normal distribution, and non-normal distributions were transformed into normal by Johnson transformation. $[\mu \pm 2\sigma]$ was used to calculate boundaries of *F* values to distinguish G4s from non-G4s group (95% prediction interval) based on the three-sigma rule of thumb.

CF factor. The global alignment analysis (30) based on Needleman-Wunsch algorithm (31) was used to search reverse complementary base pairs between X and F22. EDNAFULL (NUC4.4; <https://ftp.ncbi.nlm.nih.gov/blast/matrices/NUC.4.4>) was adopted as scoring matrix, gap penalty was at 10.0 and extension penalty was set as 0.5. CF factor was defined to quantify the complementarity between X and F22:

$$\text{CF Factor} = \frac{\text{Numbers of base pairs expected in (F22 + X duplex)}}{\text{Length (F22)}}$$

Other biophysical methods

All sequences were kept in corresponding buffers and denatured at 95°C for 5 min, and then cooled down to room temperature before use.

Isothermal differential absorbance spectrum (IDS) corresponds to the difference between the absorbance spectra obtained in the absence or in the presence of 100 mM KCl. Samples were tested at 3 μ M strand concentration in 1 mL of 10 mM lithium cacodylate pH 7.2 buffer. Absorbance spectra were recorded on a Cary 300 spectrophotometer (Agilent Technologies, France) at 25°C. Scan range was set as 500–200 nm with scan rate: 600 nm/min, baseline was corrected automatically.

Thermal differential absorbance spectrum (TDS) corresponds the difference between the absorbance spectra obtained at high (95°C) and low (25°C) temperature of the sample, tested at 3 μ M strand concentration, pre-folded in 1 mL of 100 mM KCl, 10 mM lithium cacodylate pH 7.2 buffer. Absorbance spectra were recorded at 25 and 95°C, respectively; other settings were as the same as IDS.

FRET-melting competition assay (FRET-MC): 15 μ M competitor sequences and 5 μ M F21T were pre-folded in 10K buffer (10 mM KCl, 90 mM LiCl, 10 mM lithium cacodylate, pH 7.2). Each well contained 25 μ l solution comprising 0.2 μ M F21T, 3 μ M competitor and 0.4% (v/v) DMSO, in the absence or presence of 0.4 μ M PhenDC3. FRET-MC experiments were performed in 96-well plates using a HT7900 RT-PCR instrument (Applied BioSystem), the FAM channel was used to collect the fluorescence signal. qPCR process was set as: 25°C 5 min; increasing temperature 0.5°C per minute, recording fluorescence, 140 cycles; then keeping plates at 25°C after measurements. ΔT_m was calculated as the difference between T_m of F21T with or without PhenDC3; where T_m was identified as the temperature related to $\frac{1}{2}$ fluorescence. The *S* Factor provides a normalized value (26) of the stabilization (ΔT_m) remaining

in the presence of X. *S* is close to 0 for G4-forming sample, and remains close to 1 when X is not forming a quadruplex.

Circular dichroism (CD) spectra were recorded on a J-1500 spectropolarimeter (Jasco, France). Pre-folded samples were tested at 3 μ M strand concentration in 1 ml of 100 mM KCl, 10 mM lithium cacodylate pH 7.2 buffer. The spectra were measured over the wavelength range of 200–340 nm at 25°C with a scan rate of 100 nm/min with and automatic baseline correction.

RESULTS

Principle of the assay

Several methods have been developed to characterize G4 structures *in vitro*: some of them are based on the spectral properties of G-quadruplexes (32,33), others on specific fluorescence light-up dyes such as NMM (34), Thioflavin T (35), and DASPMI (36). Both types of techniques consider only the structure of sequences of interest. We introduced the concept of competition in the FRET-MC assay (26), in which the sequence of interest X is in competition with a labeled probe. X, being added in large excess, can outcompete the labeled G4 probe for PhenDC3 binding if, and only if, X adopts a thermally stable quadruplex structure. In any other situation, X is unable to act as decoy for the specific G4 ligand, which remains bound to F21T and stabilizes it, as shown by an increase in T_m value. FRET-MC is therefore a thermal denaturation assay, and the thermal stability of the G4 structure adopted by X plays a decisive role: G4 competitors with a low T_m are unfolded before F21T starts to melt, meaning that they are seen as single-strands, leading to false negatives.

In contrast, in the isothermal competition assay developed here, the system uses two mono-labeled rather than one double-labeled fluorescence oligonucleotide, and the competition process is tested at a constant temperature (room temperature in most experiments described below, but the assay can be easily transposed to 37°C) to avoid issues related to thermal stability: what matters here is whether X is predominantly folded into a quadruplex or not at $\approx 25^\circ\text{C}$, not if its T_m is 40, 60 or 90°C.

The two mono-labeled partially complementary RNA strands, 37Q and F22, were initially used to design an isothermal assay to pick novel G4 ligands (27). In brief, potent G4 ligands stabilize the intramolecular quadruplex formed by 37Q, preventing it from hybridizing to F22. As a consequence, F22 remains single-stranded and its fluorescence is high. In contrast, when a compound has little or no affinity for the 37Q quadruplex, formation of the F22 + 37Q duplex is possible, leading to fluorescence quenching. This assay is transposable into 96-well format and allows the screening of many ligands or conditions.

In here, rather than testing a variety of compounds, we introduce F22 and 37Q to evidence G4 structures *in vitro*. As shown in Figure 1, the G4-characterization assay can be divided into three main steps:

1. An excess of the sequence of interest X (unknown competitor) is added to 37Q (Figure 1, left);
2. A well-known high affinity G4 ligand, PhenDC3, is added to the competitor-37Q mixture. PhenDC3 is there-

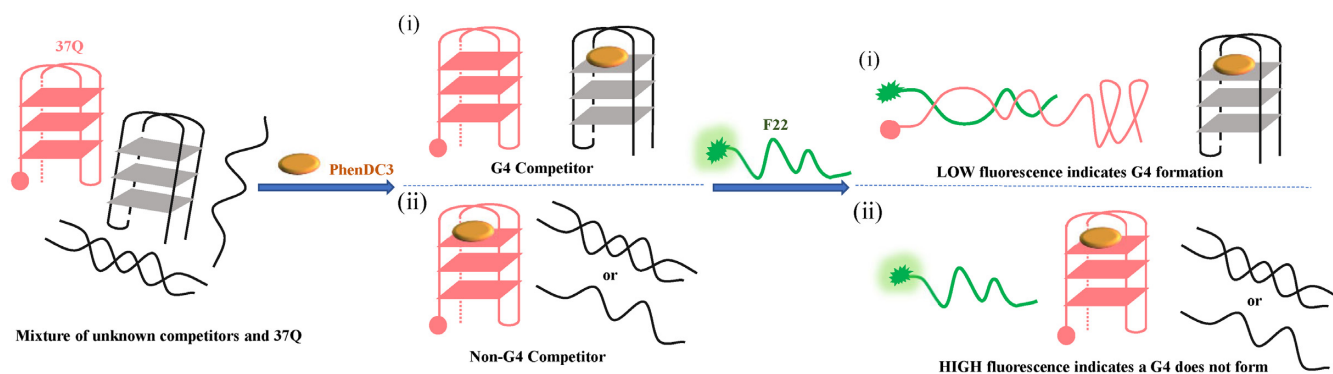


Figure 1. Principle of the iso-FRET competition assay. The modified strands (37Q, F22) are shown in color. In the first scenario (top) (i), the competitor (grey) forms a quadruplex and traps PhenDC3 (orange oval), allowing 37Q (red) to quench F22 (green) by forming a duplex. In the second scenario (bottom) (ii), the competitor does not form a quadruplex, PhenDC3 remains bound to 37Q, which cannot hybridize to F22 which emits a strong fluorescence signal.

fore ‘given the choice’ between **X** and 37Q (Figure 1, center);

3. F22 is then added, and the fluorescence signal can be recorded. Fluorescence intensity indicates whether the sequence of interest adopts or not a G4 structure (Figure 1, right).
 - i. (Figure 1, upper right part) **X** forms a quadruplex which can trap PhenDC3. As **X** is in excess (as compared to both PhenDC3 and 37Q), far less G4 ligand will be available to bind and stabilize the 37Q quadruplex, allowing duplex formation with F22, and ultimately fluorescence quenching due to fluorescein – quencher proximity.
 - ii. (Figure 1, lower right part) **X** does not form a quadruplex and does not act as a decoy for PhenDC3, which remains bound to the G4 structure formed by 37Q: formation of the F22 + 37Q duplex remains disfavored and the fluorescence emission of F22 remains high.

To validate this assay, we selected a variety of DNA and RNA sequences for which we know which structure(s) they adopt (26). This collection of 70 sequences includes a variety of quadruplex-forming motifs with various topologies, as well as single- and double-stranded DNAs and RNAs (sequences shown in Supplementary Table S1 (26)).

Practical considerations: sequential order of addition and choosing the concentrations of each component

The isothermal assay involves (at least) three different nucleic acid sequences (*i.e.* 37Q, F22 and **X**) and one G4 ligand (PhenDC3). This leaves six possible bimolecular interactions between two different partners; three of them are directly relevant for this study: (i) duplex formation between 37Q and F22, and binding between (ii) PhenDC3 and 37Q or (iii) PhenDC3 and **X**. Other possible interactions are not considered here, such as PhenDC3 binding to the F22 + 37Q duplex or to the C-rich F22 single strand: previous results have unambiguously confirmed the selectivity of PhenDC3, with little or no binding to single- and double-strands (26,28). On the other hand, possible interactions (*e.g.* partial Watson–Crick complementarity) between **X** and F22 or **X** and 37Q may generate artefacts (see below).

The key competing equilibria to be considered here are around the 37Q sequence; whether it binds to its F22 complementary sequence or to the PhenDC3 ligand, and the later event is modulated by PhenDC3 availability (whether **X** can act as decoy or not). To give a proper ‘choice’ to PhenDC3, we reasoned that adding the ligand to a well-mixed solution containing both 37Q and the competitor **X** (both were given enough time to properly fold) should favor a ‘fair’ competition for PhenDC3. Otherwise, if one of these two oligonucleotides is added after PhenDC3, re-equilibration time of the system should depend on the lifetime (k_{off}) of the quadruplex–ligand complex, for which we have little information, especially when considering a putative G4 formed by **X**.

F22 will be the last component to be added. This was previously established (27), as it is nearly impossible to reverse F22 + 37Q complex formation, as the lifetime of this duplex is extremely long at room temperature. PhenDC3 is a G4 ligand, not a duplex-destabilizing agent: one should have to first unfold the F22 + 37Q duplex to allow PhenDC3 binding to 37Q (27). The sequential addition of each component is represented in Figure 1.

We chose not to start with the addition of PhenDC3 since non-specific binding of this compound with the surface of microplate wells may be problematic. We indeed observed that artefacts may result in 96-well plates even when treated to prevent hydrophobic and ionic interactions (data not shown). This effect is abrogated by adding 1 μM salmon sperm DNA as a non-specific competitor.

As mentioned above, there are several non-covalent interactions involved in the isothermal assay, starting with the interaction of PhenDC3 with either **X** or 37Q. The principle of this experiment is that the fraction occupancy of 37Q by PhenDC3 should be significantly reduced by **X** addition if **X** adopts a quadruplex fold. The concentration of each component has to be carefully considered. As we have no *a priori* assumption of the exact K_d of PhenDC3 for the **X** quadruplex, we reasoned that **X** should be in excess as compared to 37Q and PhenDC3 in order to act as an efficient competitor. Assuming that k_{on} and k_{off} of the ligand to both structures are relatively high, what is relevant for these equilibria are the equilibrium constants (K_d). In the simpler situation where **X** does not adopt a G4 fold, PhenDC3 affinity

should be low or negligible, and one can ignore this competing equilibrium. In that case, duplex formation should be strongly inhibited, and this should happen provided that most 37Q quadruplexes are bound to PhenDC3. This can be obtained by making sure that (i) PhenDC3 is in molar excess as compared to 37Q and (ii) PhenDC3 concentration is significantly higher than the K_d for 37Q (64 nM, as determined in Supplementary Figure S1).

To summarize, the concentrations of each component should be ranked as follows:

$$[X]_0 \gg [\text{PhenDC3}]_0 > [37Q]_0 > [F22]_0$$

To study how **X** affects results, we defined 4 groups of competitors based on FRET-MC results (26), which contained strong G4s (Pu24T, cmc, 25TAG), moderately stable G4s (KRAS-22RT, SP-PGQ-1, UpsB-Q3), poor G4s (SP-PGQ-3, TBA, BmU16) and non G4s (single- or double-strands such as ds26, dT26 or Hairpin1) at different competitor concentrations in 20K buffer. The novel isothermal competition assay was used to test them. Since absolute fluorescence values are always relative (expressed in arbitrary units), they were normalized with the *F value* (26) based on the F22 fluorescence before or after F22 + 37Q hybridization; *F value* reflected the amount of PhenDC3 bound to 37Q.

As shown in Supplementary Figure S2, stable quadruplexes trap a significant fraction of PhenDC3, even at 1:1 ligand to competitor equivalents, and increasing $[X]_0$ had little effect. For duplexes and single strands, no trapping was observed at any concentration. For other quadruplexes such as KRAS-22RT, SP-PGQ-1, SP-PGQ-3, or TBA, a concentration effect was observed: raising **X** concentration led to further reduction in fluorescence intensity, until **X** concentration reached 5 μM : at this stage, *F values* were low enough and we did not investigate higher ones. Based on these observations, we chose a concentration of 5 μM for **X**, and the molar ratio between **X**, PhenDC3, 37Q and F22 was therefore fixed at 250:50:10:1.

Kinetic considerations

It is commonly believed that intramolecular G4 folding is relatively slow compared to hairpin duplex formation (37,38). Intermolecular duplex formation in the nanomolar concentration range takes a few minutes (39), while intramolecular G4 folding kinetics are impacted by a series of factors, including sequence effects (40), metal cations (41), the presence of intermediates ('dead ends') (42) and the presence of G4 ligands acting as molecular chaperones, such as PIPER (a perylene derivative) (43) and 360A (44), which have been demonstrated to accelerate G4 folding. PhenDC3 is even able to promote G4 folding within minutes in the absence of potassium (45).

To follow G4 folding under different conditions, we chose to keep the ionic strength constant, with a total concentration of mono-cations of 100 mM. G4 folding may be followed by recording absorbance at 295 nm (46). 37Q folding into G4 was followed at different K^+ concentrations, adjusting ionic strength with Li^+ . As expected, folding of 37Q in the absence of potassium (0K) proceeded extremely slowly. In the presence of K^+ , 37Q achieved near-complete

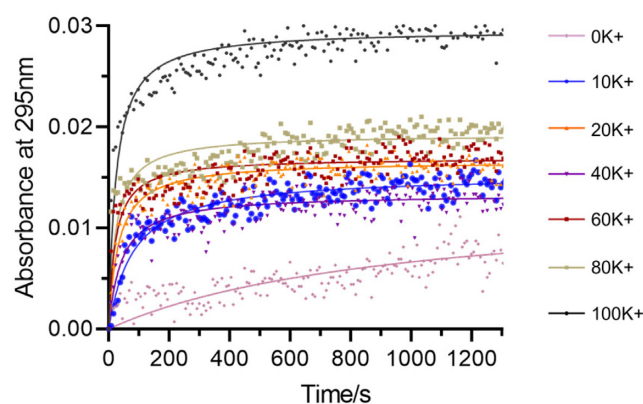


Figure 2. Kinetics of 37Q G4 folding (200 nM strand concentration) at different potassium concentrations. G4 folding was processed at RT.

folding in ≈ 5 minutes (Figure 2). Some G4 ligands such as PhenDC3 have been shown to promote G4 folding and increase k_{on} , acting as G4 chaperones (44): this is the case here, where addition of PhenDC3 shortens the G4 folding time to less than 100 s (Supplementary Figure S3).

To study F22 + 37Q duplex formation, we first considered the two oligonucleotides alone, with no competitor or PhenDC3 to simplify the model. As shown in Figure 3A, fluorescence quenching was only partial (25%) even 5 h after 37Q addition in a 100 mM K^+ buffer, while quenching reached 74% in 0K buffer. This difference can be explained by the selective stabilization of G4 structures by potassium, which partially hinders or delays F22 + 37Q hybridization, even in the absence of a G4 ligand, as previously reported (47). As expected for an intermolecular duplex, kinetics depended on strand concentration: lowering 37Q concentration delayed F22 + 37Q duplex formation (Supplementary Figure S4).

To evaluate the impact of 37Q quadruplex formation on duplex formation, we studied in parallel a control system (F22m + 37Qm), in which the F22 + 37Q pair has been mutated. F22m and 37Qm are a couple of partially complementary RNA strands also able to form a duplex involving the same number of mismatches as F22 + 37Q. However, 37Qm is unable to form a quadruplex and has no affinity for G4 ligands (27). As expected for duplex formation when no competing quadruplex is present, hybridization between F22m and 37Qm is fast, even in the nanomolar strand concentration range (39) (Figure 3B). Therefore, the relatively long (hours) equilibrium time required for proper F22 + 37Q hybridization results from quadruplex formation by 37Q, which delays, but does not prevent duplex formation.

Effects of interaction time and potassium concentration

Potassium concentration not only impacts F22 + 37Q hybridization (Figure 3A), but also the folding of G4 competitors, and the competition between 37Q and **X** for PhenDC3 binding. We initially defined the boundary between G4 **X** and non-G4 **X** roughly based on an arbitrary threshold: $F \geq 0.5$ related to non-G4s **X**, while $F < 0.5$ meant G4-forming **X**. Twelve different competitors including dif-

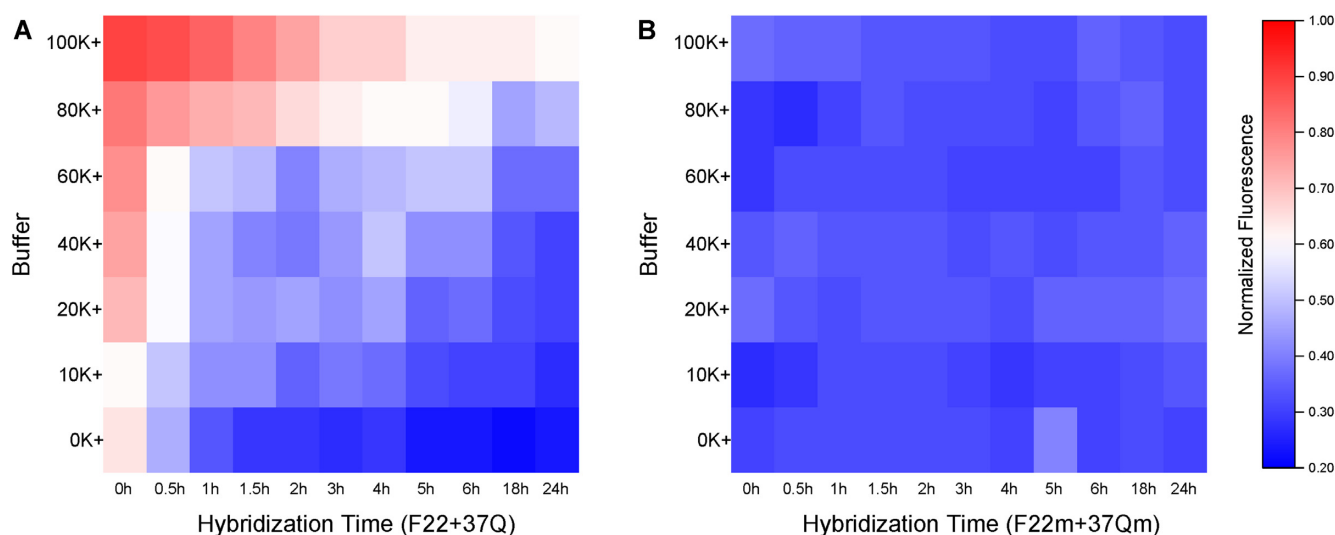


Figure 3. Time-dependent fluorescence of the F22 + 37Q system at different potassium concentrations. Concentration of F22 and F22m were 20 nM, 37Q and 37Qm were 200 nM. 37Q and 37Qm were kept in corresponding buffer for 5 min before adding F22 or F22m, respectively. Hybridizations were processed at RT. Panel A and B related to the same normalized fluorescence scale.

ferent G4s and non-G4s were taken as examples to analyze the influence of potassium. They were divided in three groups based on this rough boundary: (i) sequences in group **A** (Supplementary Figure S5A) showed high affinities to PhenDC3 under all ionic conditions and at all time points (G4 competitors, low *F* values); (ii) in contrast to group **A**, PhenDC3 remained bound to 37Q in the presence of sequences belonged to group **B** (Supplementary Figure S5B, non-G4 competitors, high *F* values); (iii) sequences in group **C** (Figure 4) are reported to fold into G4s, while their *F* values depend on specific ionic conditions and time points (poor competitors, with intermediate *F* values).

For poor competitors, *F* values went gradually up with increasing potassium concentrations: when $[K^+]$ exceeded 20 mM, *F* values were above 0.5, and this phenomenon was more apparent at short incubation times. We wanted to test whether *F* values would allow to discriminate between G4 X and non-G4 X. Although the differences between five 'good' G4 X and non-G4 X were highly significant in both 0K and 20K buffer, this difference between average *F* values (Δ Median) was actually higher in 20K (0.82) than in 0K (0.64), meaning that the discrimination between the two groups (G4 vs non-G4) is better in the presence of 20 mM potassium (20K).

$$\Delta\text{Median} = \text{Median} (F \text{ values of non-G4s}) \\ - \text{Median} (F \text{ values of G4s})$$

According to the kinetics results based on the simplified model (Figure 3A), we selected an equilibrium time of 3 h or longer as a result of the slow F22 + 37Q hybridization step. In the following experiments, we started to optimize the interaction time. An extended incubation had nearly no effect on the results obtained with groups **A** in all buffers (low *F* values), as well as for non-G4 competitors in group **B**, which kept high *F* values at all times. In contrast, SP-PGQ-3, TBA, and BmU16 in group **C** (Figure 4) exhibited a time-dependent behavior, with *F* values decreasing over

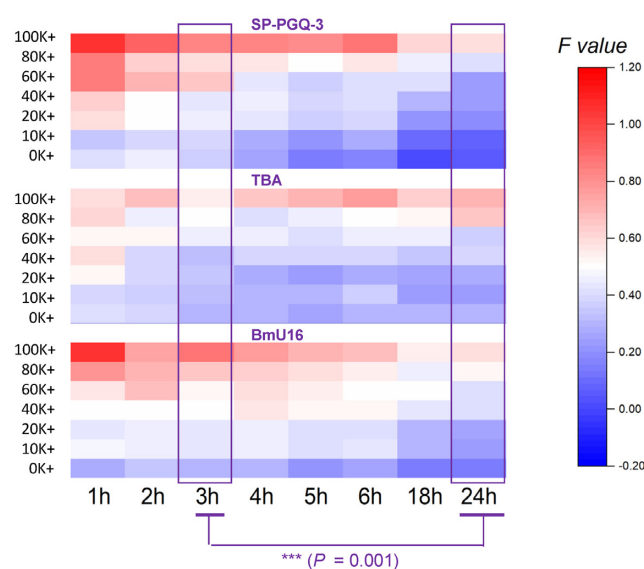


Figure 4. Effects of incubation time and ionic conditions on fluorescence intensity for group C competitors (G4-forming oligos with moderate affinity for PhenDC3). Each well contained 5 μ M competitor, 200 nM 37Q, 1 μ M PhenDC3 and 20 nM F22; control wells (F22 and F22-37Q) included 20 nM F22 in the presence or absence of 200 nM 37Q. All samples were tested in triplicate at various potassium concentrations (0–100 mM) at RT. The difference between 3 h and 24 h was performed by Wilcoxon sign test. All sub-panels related to the same *F* value scale.

time ($P = 0.001$ between 3 h and 24 h). This result illustrates the fact that even mediocre G4 competitors (G-quadruplex oligonucleotides with moderate affinities for PhenDC3 such as the thrombin binding aptamer) can be properly classified as G4-forming provided that appropriate precautions are taken. For this reason, we selected 24 h as the default interaction time in 20K buffer.

Validation of the isothermal competition assay

We employed a validated training set (26) containing a series of identified DNA and RNA G4 sequences with different topologies, as well as single- and double-stranded controls. Examples of positive controls are *cmc* and Pu24T, which both form very stable quadruplexes, while ds26 (duplex) and dT26 (single-strand) were chosen as negative controls. Each competitor oligonucleotide **X** was tested at a single concentration, and we determined the fluorescence intensity for each condition.

We divided the training sequences into 3 categories according to *F* values: (i) $F < 0.33$: G4 competitor; (ii) $0.33 \leq F < 0.54$: unknown; (iii) $F \geq 0.54$: non-G4 competitor. With these intervals, we found no outlier (false positive or negative), and all known samples behaved as predicted (Figure 5).

Regarding the temperature of the assay, while working at room temperature may be simple, homeotherms maintain body temperatures in the range of 36–42°C (48). For natural G-rich sequences found in animal genomes, it may therefore be biologically relevant to characterize them at physiological temperature. Taking *Homo sapiens* as an example, we determined if this assay would provide comparable results at 37°C. Using the same training set as previously, we found the same qualitative results: (i) $F < 0.27$: G4 competitor; (ii) $0.27 \leq F < 0.52$: unknown; (iii) $F \geq 0.52$: non-G4 competitor. (Supplementary Figure S6).

Interestingly, compared to the training set at 25°C, the difference in *F* values for the G4 and non-G4 groups was even more significant at 37°C, as a consequence of smaller standard deviations for each category. This illustrates the interest of working at physiological temperature. As a further bonus, a higher temperature accelerates F22 + 37Q hybridization, meaning that shorter incubation times may be selected.

Can poor G4 competitor be evidenced by the iso-FRET assay?

As mentioned above, the competitor **X** is added in large excess as compared to 37Q, facilitating the competition with PhenDC3. However, if **X** a forming a quadruplex with a very low affinity for PhenDC3 (as compared to 37Q), this competition may not be effective enough to give a positive result in the iso-FRET assay. We consider the assay to be ineffective when [37Q + PhenDC3] is equal or lower than [**X** + PhenDC3], meaning that the interaction between **X** and PhenDC3 is no longer dominant in the competitive binding.

Knowing the final concentrations in each well (typically 200 nM for 37Q, 5 μ M for **X**, *i.e.* a 25 \times fold excess, and 1 μ M for PhenDC3) and based on Mass action law, we can determine the equilibrium dissociation constant for the interaction between **X** and PhenDC3 (K_{dX}) that would lead to [37Q + PhenDC3] = [**X** + PhenDC3], given that we determined the dissociation constant for the interaction between a close analog of 37Q and PhenDC3 (K_{dQ}) to be 64 nM. A rapid calculation gives a value for K_{dX} of ≈ 17 μ M, a far worse (several orders of magnitude) affinity for a quadruplex that was previously determined, or measured here for PhenDC3. The affinity we determined here for Cy5-37merR

is actually a bit weaker than what has been previously reported in the literature for other quadruplexes (49,50) under different experimental conditions, with K_d as low as 2 nM. The risk of a false negative in the iso-FRET assay (a G4-forming sequence that would not act as an effective competitor) is therefore extremely low.

A potential limitation: G-rich sequences complementary to F22

The iso-FRET assay involves two probe strands, F22 and 37Q. In this part we investigate what happens when the **X** sequence is complementary to any of the probe strands:

- i. We can quickly discard the case in which **X** is Watson-Crick complementary to 37Q, as this would mean that **X** is C-rich; G4 formation would therefore be very unlikely. In any case, adding this C-rich strand would decrease the ability of the 37Q sequence to quench the fluorescence of F22 (due to hybridization to **X** in large excess), giving high *F* values indicative of non-G4 formation by **X**.
- ii. The situation where **X** is complementary to F22 is more complex, especially given that the F22 + 37Q duplex contains three point-mismatches, and one bulge. Given that the **X** oligonucleotide is in molar excess as compared to 37Q, it is clear that, if the F22 + **X** duplex is thermally more stable than F22 + 37Q, the assay will lead to an artefactual result: a FAM fluorescence signal is expected for F22 + **X** as **X** is not conjugated to a quencher. In other words, provided that **X** is able to form a stable hybrid with F22, it will appear as non G4-forming in this test no matter what is its real G4-forming propensity. This prediction was experimentally verified for a variety of G-rich probes with various levels of complementarity to F22. We defined the *CF* factor as a simple way to quantify this complementarity for each **X** and F22, and compare it to F22 + 37Q duplex.

As shown in Supplementary Table S1, 37Q held the highest *CF* Factor among all sequences in the training set, 0.77; *CF* for G4 **X** were in the range of 0.18 to 0.68, which means F22 + **X** duplex was weaker than 37Q + F22. To investigate how F22 + **X** duplex influence G4 **X**, we designed six G4 **X** (Supplementary Table S2) with a high tendency to form quadruplexes (G4Hunter scores > 1.5), as well as a good complementarity to F22 ($CF \geq 0.68$). All *CF* sequences led to significant decrease in ΔT_m of F21T in the FRET-MC assay (Supplementary Figure S7), indicating they are indeed forming G4 structures.

As shown in Figure 6, *F* values provide an excellent assessment of G4 propensity ($F < 0.35$; green horizontal line) for all sequences with *CF* factor ≤ 0.77 . Negative controls (duplexes and single strands) all give an *F* value above 0.53 (purple horizontal line) and the separation between G4 and non-G4 forming sequences is very clear. On the other hand, the assay fails to account for G4 formation with sequences having a very high level of complementarity to F22 (CF factors ≥ 0.86 ; green triangles on the right part of the figure). In other words, this assay is reliable as long as **X** is not highly complementary to F22, and this caveat can easily be anticipated when the sequence of **X** is known: one can first check

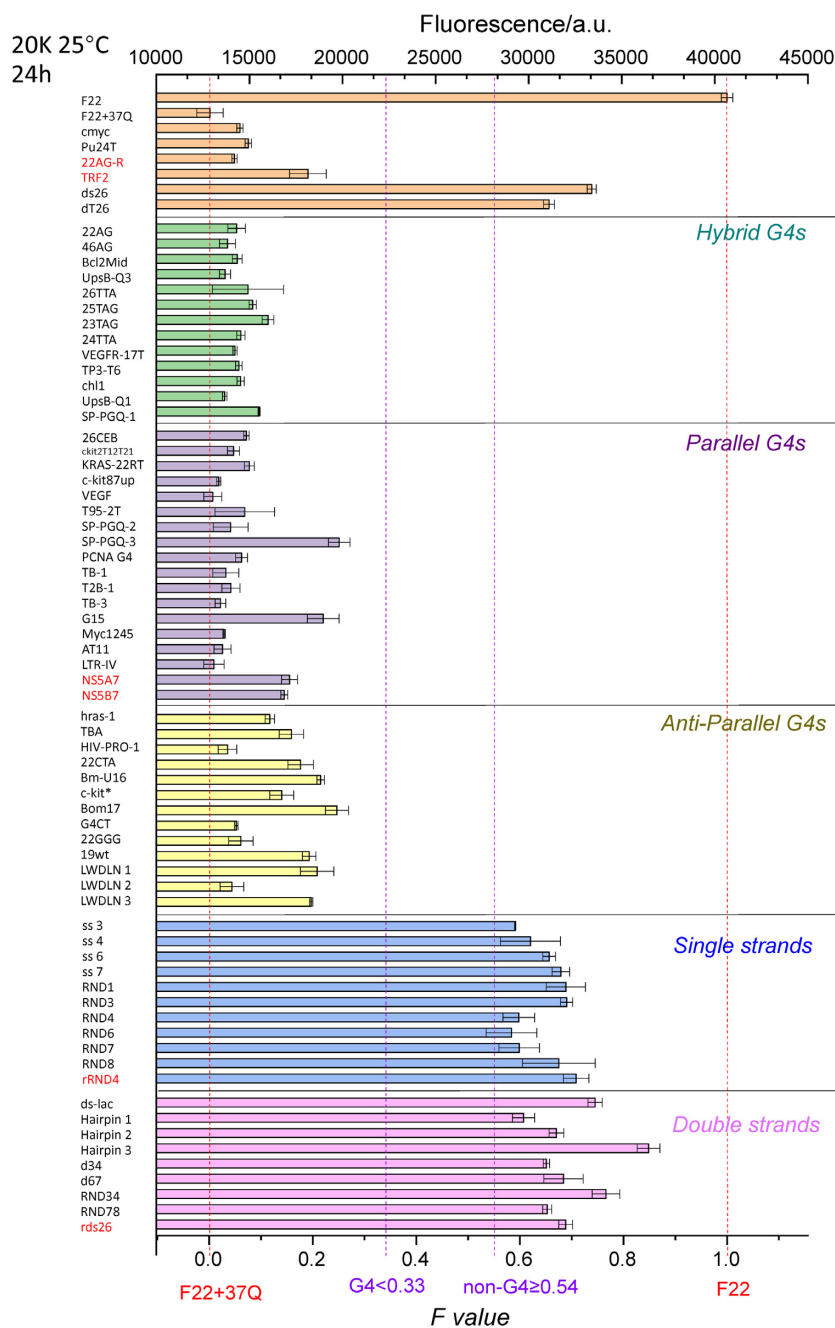


Figure 5. Fluorescence quenching in the presence of various competitors. 20 nM F22 is incubated for 24 h in the presence of 200 nM 37Q and 1 μ M PhenDC3, alone or in the presence of 5 μ M of a variety of X competitors (mostly DNA with 6 RNA samples shown in red), listed on the left. The *F* values (bottom X-axis) provides a normalized value. The four different vertical dotted lines correspond to (i) the level of fluorescence in the absence of a competitor ($F = 0$); (ii) the first threshold value at 0.33 chosen for positive samples (G4-forming sequences all exhibit *F* values between 0 and 0.33); (iii) the second threshold value at 0.54: negative controls/non G4-forming sequences all exhibit *F* values between 0.54 and 1; (iv) the level of fluorescence of F22 alone, with no 37Q added ($F = 1$). Samples were measured in 20K buffer at RT.

F22 + X complementarity and discard sequences that would form too stable hybrids.

Why using fluorescent RNA probe strands rather than DNA?

Compared to DNA oligonucleotides, RNA strands are significantly more expensive and susceptible to degradation by RNases. These inherent disadvantages prompted us to test

whether one could convert the RNA system designed by Lacroix *et al.* into a duplex-quadruplex competition assay involving oligodeoxynucleotides, dF22 and d37Q (27). As shown in Supplementary Figure S8, hybridization of dF22 to d37Q reached a plateau after two hours at room temperature, as found for F22 + 37Q hybridization. Unfortunately, quenching with the DNA oligonucleotides was not as pronounced as for the RNA system, which showed higher

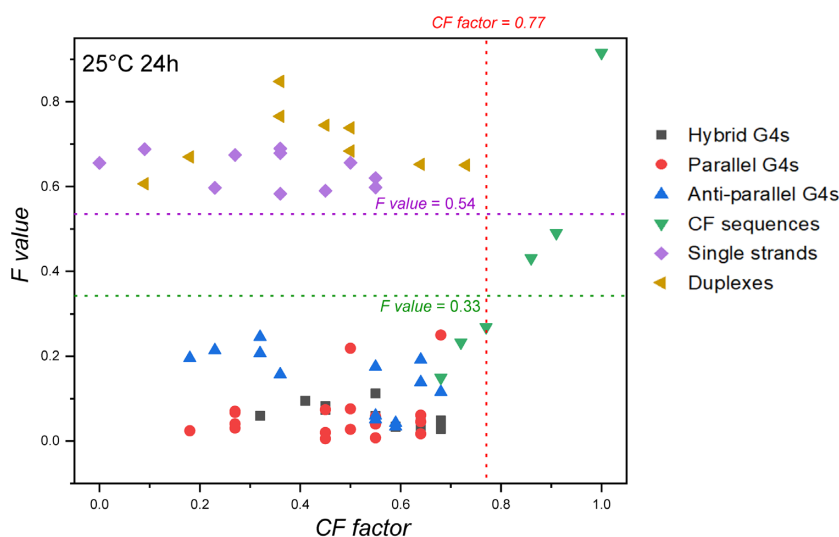


Figure 6. Fluorescence quenching in the presence of various DNA competitors. 20 nM F22 is incubated for 24 h in the presence of 200 nM 37Q and 1 μ M PhenDC3, alone or in the presence of 5 μ M of a variety of competitors. Samples were measured in 20K buffer at RT.

quenching efficiency than dF22 + d37Q under all ionic conditions at the beginning of the hybridization step (0 h). One possible reason for this difference is that the RNA to DNA substitution modifies the kinetics and thermodynamics of both the duplex and quadruplex species. For example, a RNA-RNA duplex with high G/C content has a higher k_{on} than a similar DNA duplex at 25°C (51).

We nevertheless tested whether this DNA system would be applicable to the analysis of various X sequences. As for the experiments described above, competitors from the three groups (A = G4; B = non-G4 and C = moderate G4-forming competitors) were tested at the same concentrations as for the RNA version. On the positive side, stable G4 such as cmyc and 25TAG (group A) were correctly identified as G4-competitors in this DNA-based assay, as they gave low *F values* in all settings while non-G4 forming sequences such as ds26 (group B) gave high *F values* (Supplementary Figure S9A-B), as expected in both cases. Unfortunately, Group C (modest G4 competitors) gave more erratic results with the DNA system: BmU16 showed low *F values* while both SP-PGQ-3 and TBA gave relatively high *F values*, sometimes even a bit higher than ds26 (Supplementary Figure S9C).

In other words, the DNA system worked, but not as well as the RNA one. Multiple factors may explain this difference, such as changes in duplex stability (52–54) which will also affect the stability of the F22 + X or dF22 + X duplex when X is partially complementary to F22 or dF22. As a consequence, we chose to work with the F22 + 37Q RNA system to characterize unknown sequences *in vitro*, despite the potential problems created by RNA oligonucleotides. The actual cost per point remains low, even if RNA synthesis is expensive, as the reaction volume and concentrations are low, and could be further reduced by using 384-well plates. 20 nM strand concentration for F22 in 10 μ l corresponds to 0.1 picomoles, or less than 1/10 000 of a 200 nanomole synthesis. In reality, the higher cost of RNA synthesis would start to make a difference if tens of thousands or millions of candidate sequences were to be tested. In ad-

dition, stability over time of these two labeled RNA was excellent, and stock solutions could be kept for months, if not years, with no loss of activity.

Experimental validation on viral DNA and RNA G-rich sequences

To validate the new isothermal assay, we employed twenty-three G4-prone sequences including DNAs and RNAs from pathogenic viruses (Supplementary Table S3). All of the sequences were previously predicted or demonstrated to form G4 structures. We performed classical biophysical assays to characterize their G4-forming potential before testing the isothermal assay. With the exception of HPV-16 (DNA), we found IDS and TDS spectra compatible with G4 formation for all sequences, with a negative peak around 295 nm (Supplementary Figure S10A), in agreement with FRET-MC results (Supplementary Figure S10B). CD spectra were also recorded for the DNA samples (Supplementary Figure S10C) and the proposed topologies are summarized in Supplementary Table S3 (the interpretation of CD spectra of RNA samples is more tricky, as A-form RNA duplexes give a positive peak close to the one expected for RNA parallel G4s (55)).

We analyzed these motifs with the isothermal assay at 37°C. As shown in Figure 7, isothermal assay results of all DNA competitor sequences were consistent with classical characterizations (Supplementary Figure S10): they all formed G4 structures, except HPV-16. FRET-MC of HPV-16 also gave a high *S Factor* value (>0.6), again suggesting that no G4 is formed (26) (Supplementary Figure S10). Among RNA oligomers, Nipah-NV2 gave a *F value* in the ‘undetermined’ zone ($0.27 \leq F < 0.52$), while all other methods indicated that this RNA is actually forming a G4 structure. Its relatively high *F value* could be the result of its relatively high complementarity to F22 ($CF = 0.68$; the boundary chosen for DNA should probably be lowered for RNA given the higher stability of the corresponding RNA–RNA duplex).

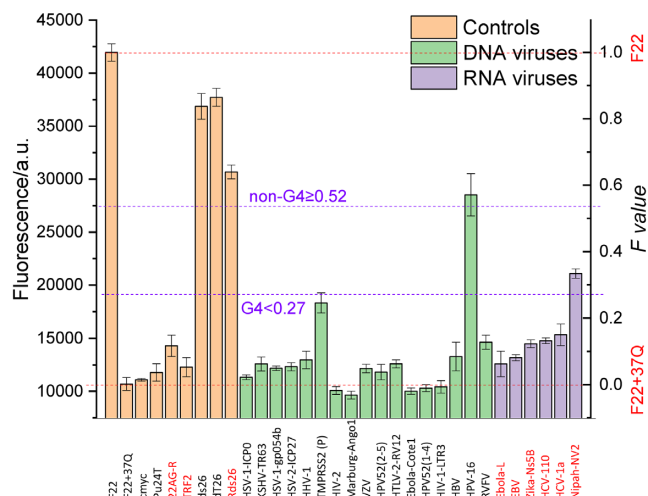


Figure 7. Fluorescence quenching in the presence of viral competitor sequences. 20 nM F22 is incubated for 24 h in the presence of 200 nM 37Q and 1 μ M PhenDC3, alone or in the presence of 5 μ M of DNA (green bars) and RNA (purple bars) virus competitors, the complementarity of each sequence to F22 is indicated by the *CF* factor value, shown above each bar. Measurements were performed in 20K buffer at 37°C. The names of RNA sequences are highlighted in red. Sequences are provided in Supplementary Table S3.

Altogether, these results illustrate that the iso-FRET assay may also be applied to RNA sequences, and that the conclusion reached by this method corroborates the analysis by other assays. One extra precaution should be taken with RNA samples though, where results can be biased by the possible complementarity between F22 and RNA X. For instance, the higher stability of the resulting RNA–RNA duplex may interfere with the isothermal assay. For this reason, we suggest to avoid testing sequences with a high complementarity (*CF* value). For example, the *F* value for Nipah-NV2 is relatively high (Figure 7), although G4 formation by this sequence was validated by other means. This ambiguous result (intermediate *F* value) may result from a high *CF* value (*CF* = 0.68, Supplementary Table S3), a relatively low affinity to PhenDC3 and/or low stability of the corresponding quadruplex.

DISCUSSION: ADVANTAGES AND LIMITATIONS OF THE ISOTHERMAL ASSAY

In this manuscript, we introduced a new method to demonstrate G4 formation *in vitro*. This isothermal assay is amenable to very high throughput. We summarize the advantages and shortcomings of this assay in Table 1, and compare its properties to the recently-developed FRET-MC assay in the next two paragraphs.

Common features between FRET-MC and iso-FRET

Both assays share a number of features: they are fluorescence-based and can be read on an inexpensive plate reader. As a consequence, many samples can be tested in parallel, and multiple positive and negative controls can be included to increase the robustness of the assay. Even if most spectroscopic instruments (e.g. a spectropolarimeter)

can be adapted using liquid handlers to improve throughput, this would correspond to a serial analysis. In contrast, FRET-MC and iso-FRET allow the testing of 96/384 samples in a parallel manner with a real time PCR machine or a fluorescence plate reader.

Both methods are relatively fast and inexpensive; low volumes (typically 25 μ l) and the concentration needed for the sample to be tested (μ M range) imply that minimal amount of the sample are required. In addition, the sequence does not need a high level of purity: the presence of minor contaminants (e.g. shorter sequences) will not perturb the assay. Consequently, expensive or time-consuming purification protocols for the X oligonucleotides are not needed. In addition, both tests should be applicable to DNA and RNA samples (as well as chemically modified nucleic acids), long sequences, and mixtures or crude oligos.

Main differences between FRET-MC and iso-FRET

Even if both assays share a number of advantages, there are significant differences between the two. First of all, iso-FRET involves one more component, as the double-labeled F21T oligo in FRET-MC is replaced by a pair of mono-labeled sequences. This makes the system slightly more complex, and the sequences to be tested should not directly interact with any of these oligonucleotides.

Iso-FRET has a potentially higher throughput, for two main reasons: (i) it is isothermal and (ii) data analysis is extremely simple, objective and direct (normalizing fluorescence intensity): there is no need to determine a T_m value and calculate a ΔT_m .

More importantly, iso-FRET can identify low-stability G4s. This can be an advantage for G4-poor genomes (some viruses such as SARS-CoV-2 contain a low density of G4-prone motifs). On the negative side, the iso-competition system is not at thermodynamic equilibrium – the result depends on the incubation time; this means that proper precautions should be taken to insure reproducibility. In addition, given that an interaction between X and F22 would perturb this assay, the sequence of X must be known; in contrast, FRET-MC can be performed blindly to characterize G-rich sequences. Finally, the fluorescence threshold in iso-FRET, chosen to distinguish between G4 and non-G4, requires a prior calibration with known controls.

CONCLUSION

The high throughput isothermal FRET assay enabled us to characterize structures of unknown G-rich sequences *in vitro*: if the unknown sequence X folds in a G4 structure, it induces fluorescence quenching by binding PhenDC3. In contrast, if X remains single-stranded or forms a different structure, the fluorescence signal remains high. This reliable so-called iso-FRET assay has been validated by several known oligonucleotide sequences forming G4 structures with different topologies, including parallel, anti-parallel, and hybrid structures, as well as double- and single-stranded DNAs and RNAs. Compared to the previous FRET-MC assay, this isothermal assay allows to process samples at 37°C, which could help indicate whether a particular sequence forms a G4 under physiological conditions.

Table 1. Comparison of FRET-MC and isothermal competition assay to characterize G4-forming sequences. Attractive features are shown in bold characters, while potential disadvantages are underlined

Parameter	FRET-MC	Iso-competition
Signal	Fluorescence (T_m)	Fluorescence (quenching)
Data analysis	T_m determination	Simple normalization
Temperature	Variable: 25–95°C	Isothermal (adjustable 20–37°C)
System	At equilibrium	<u>Not at equilibrium (under kinetic control)</u>
Number of partners	3 F21T, PhenDC3, X	<u>4 37Q, F22, PhenDC3, X</u>
Throughput	Hundreds/day	Thousands/day
Analyzed samples	DNA & RNA	DNA & RNA ^a
Ionic strength	Near-physiological	Near-physiological
Volume	25 μl	25 μl
Competitor concentration	μM range	μM range
Main Limitations/Artefacts	<u>Thermally unstable G4</u>	<u>Complementarity to F22</u>

^a Although not tested, both assays should be transposable to other nucleic acids modifications (e.g. PNA; 2'OMe), keeping in mind that, for iso-competition, complementarity to F22 may be a problem if very stable duplexes are expected to form between X and this RNA.

Importantly, iso-FRET eliminates the false negative results generated by low thermally stable G4s identified by FRET-MC. Conversely, limited by the slow hybridization rate (and slow competition process for weak G4s), iso-FRET is not a system working at thermodynamics equilibrium. In addition, the use of two mono-labelled fluorescent oligonucleotides, impedes the application of this method to G-rich sequences that show high complementary to F22 (one of probe strands). We finally applied iso-FRET to G4-prone motifs in virus genomes, and its results were confirmed by classical spectroscopic methods. The proposed isothermal competition assay constitutes a new biophysical method that can be added to the G4 toolbox required to characterize G4 structures *in vitro*.

DATA AVAILABILITY

All data is available in the supplementary information section.

SUPPLEMENTARY DATA

[Supplementary Data](#) are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Laurent Lacroix and Anne Cucchiari for helpful discussions.

FUNDING

ANR G4Access [ANR-20-CE12-0023] and ICARE [ANR-21-CE44 to J.L.M.]; Chinese Scholarship Council [201906340018 to Y.L.]. Funding for open access charge: Inserm.

Conflict of interest statement. None declared.

REFERENCES

- Agarwala, P., Pandey, S. and Maiti, S. (2015) The tale of RNA G-quadruplex. *Org. Biomol. Chem.*, **13**, 5570–5585.
- Sen, D. and Gilbert, W. (1988) Formation of parallel four-stranded complexes by guanine-rich motifs in DNA and its implications for meiosis. *Nature*, **334**, 364–366.
- Majee, P., Mishra, S.K., Pandya, N., Shankar, U., Pasadi, S., Muniyappa, K., Nayak, D. and Kumar, A. (2020) Identification and characterization of two conserved G-quadruplex forming motifs in the nipah virus genome and their interaction with G-quadruplex specific ligands. *Sci. Rep.*, **10**, 1477.
- Wang, S.R., Zhang, Q.Y., Wang, J.Q., Ge, X.Y., Song, Y.Y., Wang, Y.F., Li, X.D., Fu, B.S., Xu, G.H., Shu, B. *et al.* (2016) Chemical targeting of a G-quadruplex RNA in the ebola virus 1 gene. *Cell Chem Biol*, **23**, 1113–1122.
- Brazda, V., Luo, Y., Bartas, M., Kaura, P., Porubiakova, O., Stastny, J., Pecinka, P., Verga, D., Da Cunha, V., Takahashi, T.S. *et al.* (2020) G-Quadruplexes in the archaea domain. *Biomolecules*, **10**, 1349.
- Qin, Y. and Hurley, L.H. (2008) Structures, folding patterns, and functions of intramolecular DNA G-quadruplexes found in eukaryotic promoter regions. *Biochimie*, **90**, 1149–1171.
- Prorok, P., Artufel, M., Aze, A., Coulombe, P., Peiffer, I., Lacroix, L., Guédin, A., Mergny, J.L., Damaschke, J., Schepers, A. *et al.* (2019) Involvement of G-quadruplex regions in mammalian replication origin activity. *Nat. Commun.*, **10**, 3274.
- Paeschke, K., Bochman, M.L., Garcia, P.D., Cejka, P., Friedman, K.L., Kowalczykowski, S.C. and Zakian, V.A. (2013) Pif1 family helicases suppress genome instability at G-quadruplex motifs. *Nature*, **497**, 458–462.
- Paeschke, K., Capra, J.A. and Zakian, V.A. (2011) DNA replication through G-quadruplex motifs is promoted by the *Saccharomyces cerevisiae* pif1 DNA helicase. *Cell*, **145**, 678–691.
- Balasubramanian, S., Hurley, L.H. and Neidle, S. (2011) Targeting G-quadruplexes in gene promoters: a novel anticancer strategy? *Nat. Rev. Drug Discov.*, **10**, 261–275.
- Lago, S., Nadai, M., Cernilogar, F.M., Kazerani, M., Dominiguez Moreno, H., Schotta, G. and Richter, S.N. (2021) Promoter G-quadruplexes and transcription factors cooperate to shape the cell type-specific transcriptome. *Nat. Commun.*, **12**, 3885.
- Hansel-Hertsch, R., Beraldi, D., Lensing, S.V., Marsico, G., Zyner, K., Parry, A., Di Antonio, M., Pike, J., Kimura, H., Narita, M. *et al.* (2016) G-quadruplex structures mark human regulatory chromatin. *Nat. Genet.*, **48**, 1267–1272.
- Law, M.J., Lower, K.M., Voon, H.P., Hughes, J.R., Garrick, D., Viprakasit, V., Mitson, M., De Gobbi, M., Marra, M., Morris, A. *et al.* (2010) ATR-X syndrome protein targets tandem repeats and influences allele-specific expression in a size-dependent manner. *Cell*, **143**, 367–378.
- Valton, A.L., Hassan-Zadeh, V., Lema, I., Boggetto, N., Alberti, P., Saintome, C., Riou, J.F. and Prioleau, M.N. (2014) G4 motifs affect origin positioning and efficiency in two vertebrate replicators. *EMBO J.*, **33**, 732–746.
- Bedrat, A., Lacroix, L. and Mergny, J.L. (2016) Re-evaluation of G-quadruplex propensity with G4Hunter. *Nucleic Acids Res.*, **44**, 1746–1759.
- Toshniwal, P., Nguyen, M., Guédin, A., Viola, H., Ho, D., Kim, Y., Bhatt, U., Bond, C.S., Hool, L., Hurley, L.H. *et al.* (2019) TGF-beta-induced fibrotic stress increases G-quadruplex formation in human fibroblasts. *FEBS Lett.*, **593**, 3149–3161.

17. Gazanion, E., Lacroix, L., Alberti, P., Gurung, P., Wein, S., Cheng, M., Mergny, J.L., Gomes, A.R. and Lopez-Rubio, J.J. (2020) Genome wide distribution of G-quadruplexes and their impact on gene expression in malaria parasites. *PLoS Genet.*, **16**, e1008917.
18. Saad, M., Guédin, A., Amor, S., Bedrat, A., Tourasse, N.J., Fayyad-Kazan, H., Pratiel, G., Lacroix, L. and Mergny, J.L. (2019) Mapping and characterization of G-quadruplexes in the genome of the social amoeba *dictyostelium discoideum*. *Nucleic Acids Res.*, **47**, 4363–4374.
19. Bohalova, N., Cantara, A., Bartas, M., Kaura, P., Stastny, J., Pecinka, P., Fojta, M., Mergny, J.L. and Brazda, V. (2021) Analyses of viral genomes for G-quadruplex forming sequences reveal their correlation with the type of infection. *Biochimie*, **186**, 13–27.
20. Brazda, V., Porubiakova, O., Cantara, A., Bohalova, N., Coufal, J., Bartas, M., Fojta, M. and Mergny, J.L. (2021) G-quadruplexes in H1N1 influenza genomes. *BMC Genomics*, **22**, 77.
21. Brazda, V., Kolomaznik, J., Lysek, J., Bartas, M., Fojta, M., Stastny, J. and Mergny, J.L. (2019) G4Hunter web application: a web server for G-quadruplex prediction. *Bioinformatics*, **35**, 3493–3495.
22. Huppert, J.L. and Balasubramanian, S. (2005) Prevalence of quadruplexes in the human genome. *Nucleic Acids Res.*, **33**, 2908–2916.
23. Todd, A.K., Johnston, M. and Neidle, S. (2005) Highly prevalent putative quadruplex sequence motifs in human DNA. *Nucleic Acids Res.*, **33**, 2901–2907.
24. Tu, J., Duan, M., Liu, W., Lu, N., Zhou, Y., Sun, X. and Lu, Z. (2021) Direct genome-wide identification of G-quadruplex structures by whole-genome resequencing. *Nat. Commun.*, **12**, 6014.
25. Adrian, M., Heddi, B. and Phan, A.T. (2012) NMR spectroscopy of G-quadruplexes. *Methods*, **57**, 11–24.
26. Luo, Y., Granzhan, A., Verga, D. and Mergny, J.-L. (2021) FRET-MC: a fluorescence melting competition assay for studying G4 structures in vitro. *Biopolymers*, **112**, e23415.
27. Lacroix, L., Seosse, A. and Mergny, J.L. (2011) Fluorescence-based duplex-quadruplex competition test to screen for telomerase RNA quadruplex ligands. *Nucleic Acids Res.*, **39**, e21.
28. De Cian, A., DeLemos, E., Mergny, J.-L., Teulade-Fichou, M.-P. and Monchaud, D. (2007) Highly efficient G-quadruplex recognition by bisquinolinium compounds. *J. Am. Chem. Soc.*, **129**, 1856–1857.
29. Le, D.D., Di Antonio, M., Chan, L.K. and Balasubramanian, S. (2015) G-quadruplex ligands exhibit differential G-tetrad selectivity. *Chem. Commun. (Camb.)*, **51**, 8048–8050.
30. Madeira, F., Park, Y.M., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., Basutkar, P., Tivey, A.R.N., Potter, S.C., Finn, R.D. *et al.* (2019) The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res.*, **47**, W636–W641.
31. Needleman, S.B. and Wunsch, C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443–453.
32. Del Villar-Guerra, R., Trent, J.O. and Chaires, J.B. (2018) G-Quadruplex secondary structure obtained from circular dichroism spectroscopy. *Angew. Chem. Int. Ed Engl.*, **57**, 7171–7175.
33. Mergny, J.L., Li, J., Lacroix, L., Amrane, S. and Chaires, J.B. (2005) Thermal difference spectra: a specific signature for nucleic acid structures. *Nucleic Acids Res.*, **33**, e138.
34. Arthanari, H., Basu, S., Kawano, T.L. and Bolton, P.H. (1998) Fluorescent dyes specific for quadruplex DNA. *Nucleic Acids Res.*, **26**, 3724–3728.
35. Renaud de la Faverie, A., Guédin, A., Bedrat, A., Yatsunyk, L.A. and Mergny, J.L. (2014) Thioflavin T as a fluorescence light-up probe for G4 formation. *Nucleic Acids Res.*, **42**, e65.
36. Xie, X., Zuffo, M., Teulade-Fichou, M.P. and Granzhan, A. (2019) Identification of optimal fluorescent probes for G-quadruplex nucleic acids through systematic exploration of mono- and distyryl dye libraries. *Beilstein J. Org. Chem.*, **15**, 1872–1889.
37. Lane, A.N., Chaires, J.B., Gray, R.D. and Trent, J.O. (2008) Stability and kinetics of G-quadruplex structures. *Nucleic Acids Res.*, **36**, 5482–5515.
38. Gray, R.D., Trent, J.O., Arumugam, S. and Chaires, J.B. (2019) Folding landscape of a parallel G-quadruplex. *J. Phys. Chem. Lett.*, **10**, 1146–1151.
39. Xu, S., Zhan, J., Man, B., Jiang, S., Yue, W., Gao, S., Guo, C., Liu, H., Li, Z., Wang, J. *et al.* (2017) Real-time reliable determination of binding kinetics of DNA hybridization using a multi-channel graphene biosensor. *Nat. Commun.*, **8**, 14902.
40. Nguyen, T.Q.N., Lim, K.W. and Phan, A.T. (2020) Folding kinetics of G-quadruplexes: duplex stem loops drive and accelerate G-quadruplex folding. *J. Phys. Chem. B*, **124**, 5122–5130.
41. Bhattacharyya, D., Mirihana Arachchilage, G. and Basu, S. (2016) Metal cations in G-quadruplex folding and stability. *Front Chem*, **4**, 38.
42. Harkness, R.W., Hennecker, C., Grun, J.T., Blumler, A., Heckel, A., Schwalbe, H. and Mittermaier, A.K. (2021) Parallel reaction pathways accelerate folding of a guanine quadruplex. *Nucleic Acids Res.*, **49**, 1247–1262.
43. Han, H., Cliff, C.L. and Hurley, L.H. (1999) Accelerated assembly of G-Quadruplex structures by a small molecule. *Biochemistry*, **38**, 6981–6986.
44. De Cian, A. and Mergny, J.L. (2007) Quadruplex ligands may act as molecular chaperones for tetramolecular quadruplex formation. *Nucleic Acids Res.*, **35**, 2483–2493.
45. Aznauryan, M., Noer, S.L., Pedersen, C.W., Mergny, J.-L., Teulade-Fichou, M.-P. and Birkedal, V. (2021) Ligand binding to dynamically populated G-quadruplex DNA. *ChemBioChem*, **22**, 1811–1817.
46. Mergny, J.-L., Phan, A.-T. and Lacroix, L. (1998) Following G-quartet formation by UV-spectroscopy. *FEBS Lett.*, **435**, 74–78.
47. Alberti, P. and Mergny, J.-L. (2003) DNA duplex–quadruplex exchange as the basis for a nanomolecular machine. *Proc. Natl. Acad. Sci. U.S.A.*, **100**, 1569.
48. Ivanov, K.P. (2006) The development of the concepts of homeothermy and thermoregulation. *J. Therm. Biol.*, **31**, 24–29.
49. Bonnat, L., Bar, L., Gennaro, B., Bonnet, H., Jarjayes, O., Thomas, F., Dejeu, J., Defrancq, E. and Lavergne, T. (2017) Template-Mediated stabilization of a DNA G-Quadruplex formed in the HIV-1 promoter and comparative binding studies. *Chemistry*, **23**, 5602–5613.
50. Bonnat, L., Dautriche, M., Saidi, T., Revol-Cavalier, J., Dejeu, J., Defrancq, E. and Lavergne, T. (2019) Scaffold stabilization of a G-triplex and study of its interactions with G-quadruplex targeting ligands. *Org. Biomol. Chem.*, **17**, 8726–8736.
51. Rauzan, B., McMichael, E., Cave, R., Sevcik, L.R., Ostrosky, K., Whitman, E., Stegemann, R., Sinclair, A.L., Serra, M.J. and Deckert, A.A. (2013) Kinetics and thermodynamics of DNA, RNA, and hybrid duplex formation. *Biochemistry*, **52**, 765–772.
52. Cheatham, T.E. and Kollman, P.A. (1997) Molecular dynamics simulations highlight the structural differences among DNA:DNA, RNA:RNA, and DNA:RNA hybrid duplexes. *J. Am. Chem. Soc.*, **119**, 4805–4825.
53. Lesnik, E.A. and Freier, S.M. (1995) Relative thermodynamic stability of DNA, RNA, and DNA:RNA hybrid duplexes: relationship with base composition and structure. *Biochemistry*, **34**, 10807–10815.
54. Gyi, J.I., Lane, A.N., Conn, G.L. and Brown, T. (1998) The orientation and dynamics of the C2'-OH and hydration of RNA and DNA:RNA hybrids. *Nucleic Acids Res.*, **26**, 3104–3110.
55. Weldon, C., Eperon, I.C. and Dominguez, C. (2016) Do we know whether potential G-quadruplexes actually form in long functional RNA molecules? *Biochem. Soc. Trans.*, **44**, 1761–1768.

Iso-FRET: An isothermal competition assay to analyze quadruplex formation *in vitro*

Yu Luo^{1,2}, Daniela Verga^{2,3} & Jean-Louis Mergny¹

1. Laboratoire d'Optique et Biosciences, Ecole Polytechnique, CNRS, Inserm, Institut Polytechnique de Paris, 91128 Palaiseau, France.
2. CNRS UMR9187, INSERM U1196, Université Paris-Saclay, F-91405 Orsay, France.
3. CNRS UMR9187, INSERM U1196, Institut Curie, PSL Research University, F-91405 Orsay, France.

Supplementary information

Content:

Titles	Content	Pages
Table S1	Training set of DNA and RNA sequences.	<i>ii</i>
Table S2	DNA G-rich sequences reverse complemented to F22.	<i>v</i>
Table S3	G-rich sequences in viruses.	<i>vi</i>
Figure S1	Dissociation constant (Kd) of cy5-37merR for PhenDC3.	<i>vii</i>
Figure S2	F22 quenching by 37Q and PhenDC3, alone or in the presence of known competitors (X) at different concentrations (1-10 μ M) after 24 h.	<i>viii</i>
Figure S3	Kinetics of 200 nM 37Q G4 folding at different potassium concentrations in the presence of 1 μ M PhenDC3.	<i>ix</i>
Figure S4	Time-dependent fluorescence of the F22+37Q system at different quencher concentrations.	<i>x</i>
Figure S5	Effects of G4 or non-G4 competitors on F22+37Q fluorescence intensity, tested at different times and potassium concentrations.	<i>xi</i>
Figure S6	Fluorescence quenching in the presence of various competitors at 37 °C.	<i>xii</i>
Figure S7	FRET-MC results of CF series competitors.	<i>xiii</i>
Figure S8	Time-dependent fluorescence of the system involving fluorescently-labeled DNA oligonucleotides (dF22+d37Q) at different potassium concentrations.	<i>xiv</i>
Figure S9	Effects of different competitors on dF22+d37Q fluorescence intensity, tested at different times and potassium concentrations.	<i>xv</i>
Figure S10	IDS, TDS, CD spectrum and FRET-MC results of sequences in viruses.	<i>xvi</i>
References	34 Additional References	<i>xix</i>

Table S1. Training set of DNA and RNA sequences.

Acronym	Sequence (5'-3')	Reported conformation	Reference / PDB	CF Factor
F22	FAM-UGGCCCCGUUCGCCCCUCCCGGG	-	-	
F22m	FAM-UGGCCCCGUUCGCUUCUCUCGGG	-	-	-
37Q	GGGUUGCGGAGGGUGGGCCUGGGAGGGGUG GUGGCCA-BHQ1	-	-	0.77
37Qm	GAGUUGCGAAGAGUGAGCCUGAGAGAAGUG AUGGCCA-BHQ1	-	-	-
Cy5-37merR	Cy5- GGGUUGCGGAGGGUGGGCCUGGGAGGGG UGGUGGCCA	-	-	-
dF22	FAM-TGGCCCGTTCGCCCTCCCGGG	-	-	-
d37Q	GGGTTGCGGAGGGTGGGCCTGGGAGGGGTGG TGGCCA-BHQ1	-	-	0.77
22AG	AGGGTTAGGGTTAGGGTTAGGG	hybrid DNA G4	[1]	0.55
46AG	AGGGTTAGGGTTAGGGTTAGGGTTAGGGTTA GGGTTAGGGTTAGGG	hybrid DNA G4	[1]	0.59
Bcl2Mid	GGGCGCGGAGGAATTGGGCGGG	hybrid DNA G4	2F8U	0.64
UpsB-Q3	CAGGGTTAAGGGTATACATTTAGGGGTTAGG GTT	hybrid DNA G4	[2]	0.64
26TTA	TTAGGGTTAGGGTTAGGGTTAGGGTT	hybrid DNA G4	2JPZ	0.45
25TAG	TAGGGTTAGGGTTAGGGTTAGGGTT	hybrid DNA G4	2JSL	0.45
23TAG	TAGGGTTAGGGTTAGGGTTAGGG	hybrid DNA G4	2JSK	0.55
24TTA	TTAGGGTTAGGGTTAGGGTTAGGG	hybrid DNA G4	2JSL	0.32
VEGFR-17T	GGGTACCCGGGTGAGGTGCGGGGT	hybrid DNA G4	5ZEV	0.68
TP3-T6	TGGGGTCCGAGGCGGGCTTGGG	hybrid DNA G4	6AC7	0.55
chl1	GGGTGGGGAAGGGGTGGGT	hybrid DNA G4	2KPR	0.55
UpsB-Q1	CAGGGTTAAGGGTATAACTTTAGGGGTTAGG GTT	hybrid DNA G4	5MTA	0.68
SP-PGQ-1	GGGCAACTTGGCTGGGGTCTAGTTCCACGGG ACGGG	hybrid DNA G4	[3]	0.41
26CEB	AAGGGTGGGTGTAAGTGTGGGTGGGT	parallel DNA G4	2LPW	0.27
c-kit2-T12T21	CGGGCGGGCGCTAGGGAGGGT	parallel DNA G4	2KYP	0.64
KRAS-22RT	AGGGCGGTGTGGGAATAGGGAA	parallel DNA G4	5I2V	0.50
Pu24T	TGAGGGTGGTGAGGGTGGGGAAGG	parallel DNA G4	2A5P	0.45
c-kit87up	AGGGAGGGCGCTGGGAGGAGGG	parallel DNA G4	2O3M	0.64
VEGF	CGGGGCGGGCCTTGGGCGGGGT	parallel DNA G4	2M27	0.45
cmyc	TGAGGGTGGGTAGGGTGGGTAA	parallel DNA G4	1XAV	0.55
T95-2T	TTGGGTGGGTGGGTGGGT	parallel DNA G4	2LK7	0.27
SP-PGQ-2	GGGCTAGTGGGGGAGGGGG	parallel DNA G4	[3]	0.55
SP-PGQ-3	GGGCTAATAGGGAGAGCAGGGACGGGG	parallel DNA G4	[3]	0.68
PCNA DNA G4	CAGGGCGACGGGGGCGGGGCGGGGCG	parallel DNA G4	[4]	0.64

TB-1	TTGTGGTGGGTGGGTGGGT	parallel DNA G4	2M4P	0.27
T2B-1	TTGTTGGTGGGTGGGTGGGT	parallel DNA G4	[5]	0.27
TB-3	TTGGGTGTGGTGGGTGGGT	parallel DNA G4	[5]	0.45
G15	TTGGGGGGGGGGGGGGGGT	parallel DNA G4	2MB2	0.50
Myc1245	TTGGGGAGGGTTTTTAAGGGTGGGGAAT	parallel DNA G4	6NEB	0.50
AT11	TGGTGGTGGTTGTTGTGGTGGTGGTGGT	parallel DNA G4	2N3M	0.18
LTR-IV	CTGGGCGGGACTGGGGAGTGGT	parallel DNA G4	2N4Y	0.55
hras-1	TCGGGTTGCGGGCGCAGGGCACGGGCG	anti-parallel DNA G4	[6]	0.68
TBA	GGTTGGTGTGGTTGG	anti-parallel DNA G4	148D	0.36
HIV-PRO-1	TGGCCTGGGCGGGACTGGG	anti-parallel DNA G4	[7]	0.59
22CTA	AGGGCTAGGGCTAGGGCTAGGG	anti-parallel DNA G4	[8]	0.55
Bm-U16	TAGGTTAGGTTAGGTUAGG	anti-parallel DNA G4	[9]	0.23
c-kit*	GGCGAGGAGGGGCGTGGCCGGC	anti-parallel DNA G4	6GH0	0.64
Bom17	GGTTAGGTTAGGTTAGG	anti-parallel DNA G4	[10]	0.32
DNA G4CT	GGGGCTGGGGCTGGGGCTGGGG	anti-parallel DNA G4	[11]	0.55
22GGG	GGGTTAGGGTTAGGGTTAGGGT	anti-parallel DNA G4	2KF8	0.55
19wt	GGGGGAGGGGTACAGGGGTACAGGGG	anti-parallel DNA G4	6FTU	0.64
LWDLN 1	GGGTTTGGGTTTTGGGAGGG	anti-parallel DNA G4	5J05	0.32
LWDLN 2	GGGGTTGGGGTTTTGGGGAAGGGG	anti-parallel DNA G4	2M6W	0.59
LWDLN 3	GGTTTGGTTTTGGTTGG	anti-parallel DNA G4	5J4W	0.18
ss 3	GTCGCCGGGCCAGTCGTCCATAC	ssDNA	-	0.45
ss 4	GTATGGACGACTGGCCCGGCGAC	ssDNA	-	0.55
ss 6	GACGTGTCGAAAGAGCTCCGATTA	ssDNA	-	0.50
ss 7	TAATCGGAGCTCTTTCGACACGTC	ssDNA	-	0.36
RND1	CTATACGAAAACCTTTTGTATCATT	ssDNA	-	0.09
RND2	AATGATACAAAAGGTTTTCGTATAG	ssDNA	-	0.23
RND3	TAACGTTTATAATGTAGTCTCATTA	ssDNA	-	0.36
RND4	TAATGAGACTACATTATAAACGTTA	ssDNA	-	0.23
RND6	GTTGTCATTGCCCGCAATAATTCT	ssDNA	-	0.36
RND7	GCCTTGCGGAGGCATGCGTCATGCT	ssDNA	-	0.55
RND8	AGCATGACGCATGCCTCCGCAAGGC	ssDNA	-	0.27
dT26	TTTTTTTTTTTTTTTTTTTTTTTTTTT	ssDNA	-	0.00
ds26	CAATCGGATCGAATTTCGATCCGATTG	dsDNA	-	0.50
ds-lac	GAATTGTGAGCGCTCACAAATTC	dsDNA	-	0.45

Hairpin 1	GGATTCTTGGATTTTCCAAGAATCC	dsDNA	-	0.09
Hairpin 2	TCGGTATTGTGTTTCACAATACCGA	dsDNA	-	0.18
Hairpin 3	AGGACGGTGTATTTTACACCGTCCT	dsDNA	-	0.36
d34	GTCGCCGGGCCAGTCGTCCATAC	dsDNA	-	0.73
	GTATGGACGACTGGCCCGGCGAC			
d67	GACGTGTCGAAAGAGCTCCGATTA	dsDNA	-	0.50
	TAATCGGAGCTCTTTCGACACGTC			
RND34	TAACGTTTATAATGTAGTCTCATTA	dsDNA	-	0.36
	TAATGAGACTACATTATAAACGTTA			
RND78	AGAATTATTCGGGGGCAATGACAAC	dsDNA	-	0.64
	GTTGTCATTGCCCCGAATAATTCT			
22AG-R	r-AGGGUUAGGGUUAGGGUUAGGG	parallel RNA G4	[12]	0.44
TRF2	r-CGGGAGGGCGGGGAGGGC	parallel RNA G4	[13]	0.68
NS5A7	r-GUGGAGGUGGGACGGGAG	parallel RNA G4	[14]	0.64
NS5B7	r-UCGGAUGUGGCAGAGGGGGCUGGAG	parallel RNA G4	[14]	0.54
rRND4	r-UAAUGAGACUACAUUAUAAACGUUA	ss RNA	-	0.23
rds26	r-CAAUCGGAUCGAAUUCGAUCCGAUUG	ds RNA	-	0.5

Table S2. DNA G-rich sequences reverse complemented to F22. *

Chemistry	Name	Sequence (5'-3')	G4Hunter Score	CF Factor
DNA	CF1	AAGGGAGGGG <u>CCCGGGAGGGGCGAA</u>	1.64	0.68
DNA	CF2	AGGGAGGGG <u>CCCGGGAGGGGCGAAC</u>	1.6	0.72
DNA	CF3	GGGAGGGG <u>CCCGGGAGGGGCGAACG</u>	1.64	0.77
DNA	CF4	AAGGGG <u>CCCGGGAGGGGCGAACGGG</u>	1.6	0.86
DNA	CF5	AGGGG <u>CCCGGGAGGGGCGAACGGGC</u>	1.56	0.91
DNA	CF6	GGG <u>CCCGGGAGGGGCGAACGGGCCA</u>	1.59	1.00

* Bases underlined & in bold characters can be paired to F22.

Table S3. G-rich sequences in viruses.

Chem	Acronym	Sequence (5'-3')	Ref.	Conformation	G4Hunter Scores	CF Factor
DNA	HSV-1-ICP0	GGGGAGGGGAAAGGCGTGGGG	[15]	Hybrid G4	2.48	0.59
DNA	KSHV-TR63	GGGGCGGGGACGGGGGAGGGG	[16]	Hybrid G4	3.18	0.55
DNA	HSV-1-gp054b	GGGGTTGGGGTTGGGGTTGGGG	[17]	Hybrid G4	2.91	0.32
DNA	HSV-2-ICP27	GGGGACGGCGGGGGCGGGGG	[15]	Parallel G4	2.85	0.59
DNA	HHV-1	GGGGGGTGTGTTTTGGGGGGG	[18]	Parallel G4	2.57	0.50
DNA	TMPRSS2 (P) ^a	GGAGGGCGGCCGGGGGCGGGGGCGGGCGGG G	[19]	Parallel G4	2.41	0.59
DNA	HIV-2 ^b	TGGGGGGAGGACATGGGCCGGGAGGGT	[20]	Parallel G4	2.00	0.36
DNA	Marburg-Ango1 ^b	AGAGGGGACTGGTTGGGGTCTGGGTGGTA	[21]	Hybrid G4	1.96	0.5
DNA	VZV	GGGCGGGCGACGGGCGGG	[15]	Parallel G4	1.83	0.59
DNA	HPV52(2-5)	GGGCAGGGGACACAGGGTAGGG	[22]	Hybrid G4	1.82	0.45
DNA	HTLV-2-RV12 ^b	GGGGAAGTGGGTAAGGGTGAGG	[23]	Hybrid G4	1.82	0.55
DNA	Ebola-Cote1 ^b	TGGGGTGGTGTTTGAGGGTTTGGGT	[21]	Hybrid G4	1.74	0.45
DNA	HPV52(1-4)	GGGTAGGGCAGGGACACAGGGT	[24]	Hybrid G4	1.74	0.59
DNA	HIV-1-LTR3 ^b	GGGAGGCGTGGCCTGGGCGGGACTGGGG	[25]	Hybrid G4	1.43	0.59
DNA	HBV	GGGAGTGGGAGCATTCTGGGCCAGGG	[26]	Hybrid G4	1.28	0.77
DNA	HPV-16	AGGGTCGGGTACAGGCGGACGCACTGGGT	[27]	non-G4	1.11	0.23
DNA	RVFV ^b	GGGGTTGGGGGGAAGGGGAGTTGGGG	[28]	Mix G4 topologies	2.85	0.64
RNA	Ebola-L	r-GGGGUCAUAUGGGAGGGAUUGAAGG	[29]	G4	1.52	0.14
RNA	EBV	r-GGGGCAGGAGCAGGAGGA	[30]	G4	1.5	0.50
RNA	Zika-Ns5B	r-GGAUGUGGCAGAGGGGGCUGGAG	[31]	G4	1.43	0.55
RNA	HCV-110	r-GGGAGGGGGGUCCUGGAGG	[32]	G4	2.05	0.55
RNA	HCV-1a	r-GGGCUGCGGGUGGGCGGGA	[33]	G4	1.79	0.50
RNA	Nipah-NV2	r-GUGCGGGGAGGUAAAGAGGAGGCCAGG	[34]	G4	1.11	0.68

^a *TMPRSS2* in the genome of influenza A virus.

^b G-rich motifs are in cDNA of retroviruses.

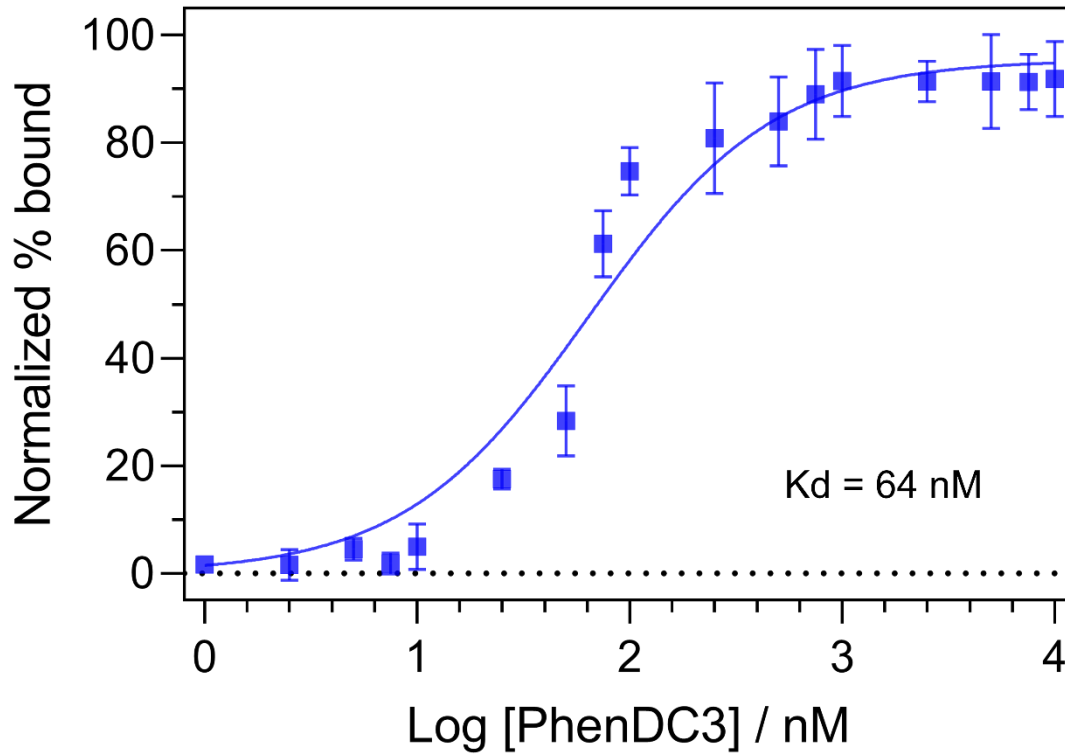


Figure S1. Dissociation constant (K_d) determination from the binding curve obtained by titration of 10 nM cy5-37merR with increasing concentration of PhenDC3 in 20K buffer. The K_d was fitted by one-site binding model, R-square = 0.97.

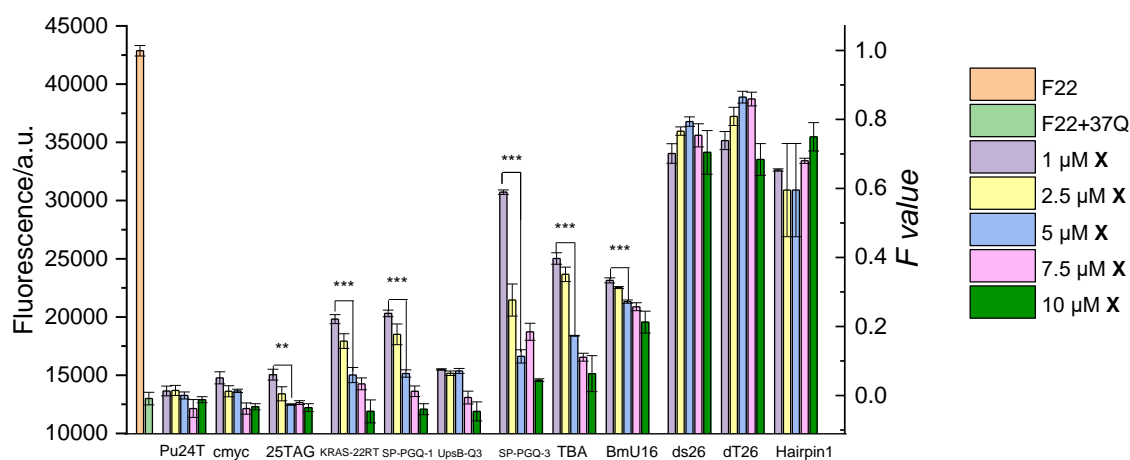


Figure S2. 20 nM F22 quenching by 200 nM 37Q and 1 μ M PhenDC3, alone or in the presence of known competitors (X) at different concentrations (1-10 μ M) after 24 h. The *S Factor* (right Y-axis) provides a normalized value. Samples were measured in 20K buffer at RT. *F values* difference between 1 μ M and 5 μ M competitor concentrations were determined by *t*-test, ** $p < 0.01$; *** $p < 0.001$.

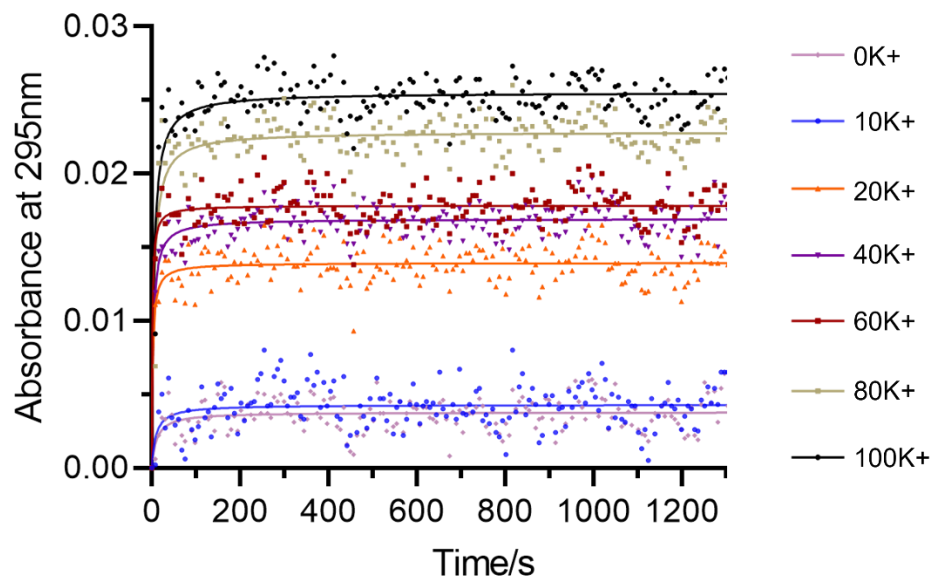


Figure S3. Kinetics of 37Q G4 folding (200 nM strand concentration) at different potassium concentrations in the presence of 1 μ M PhenDC3. G4 folding was processed at RT.

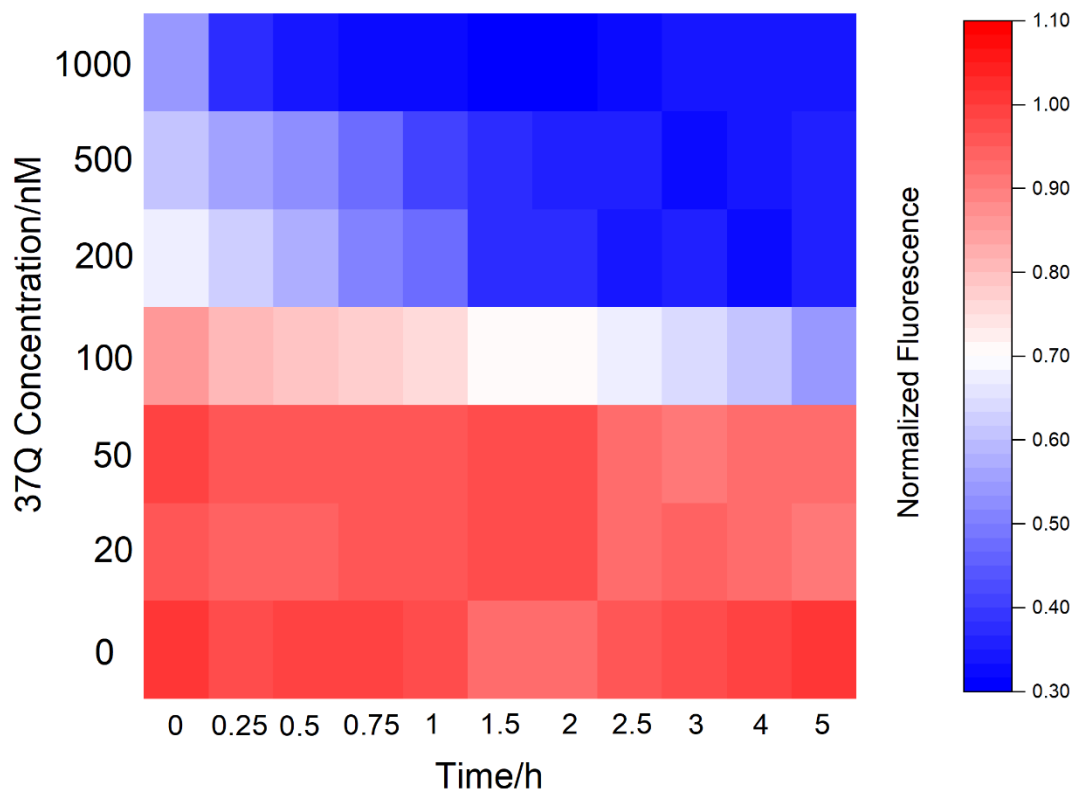


Figure S4. Time-dependent fluorescence of the F22+37Q system at different quencher concentrations. 37Q at desired concentration was kept in corresponding buffer for 5 min before adding F22 at 20 nM. Hybridizations were processed in 20K buffer at RT.

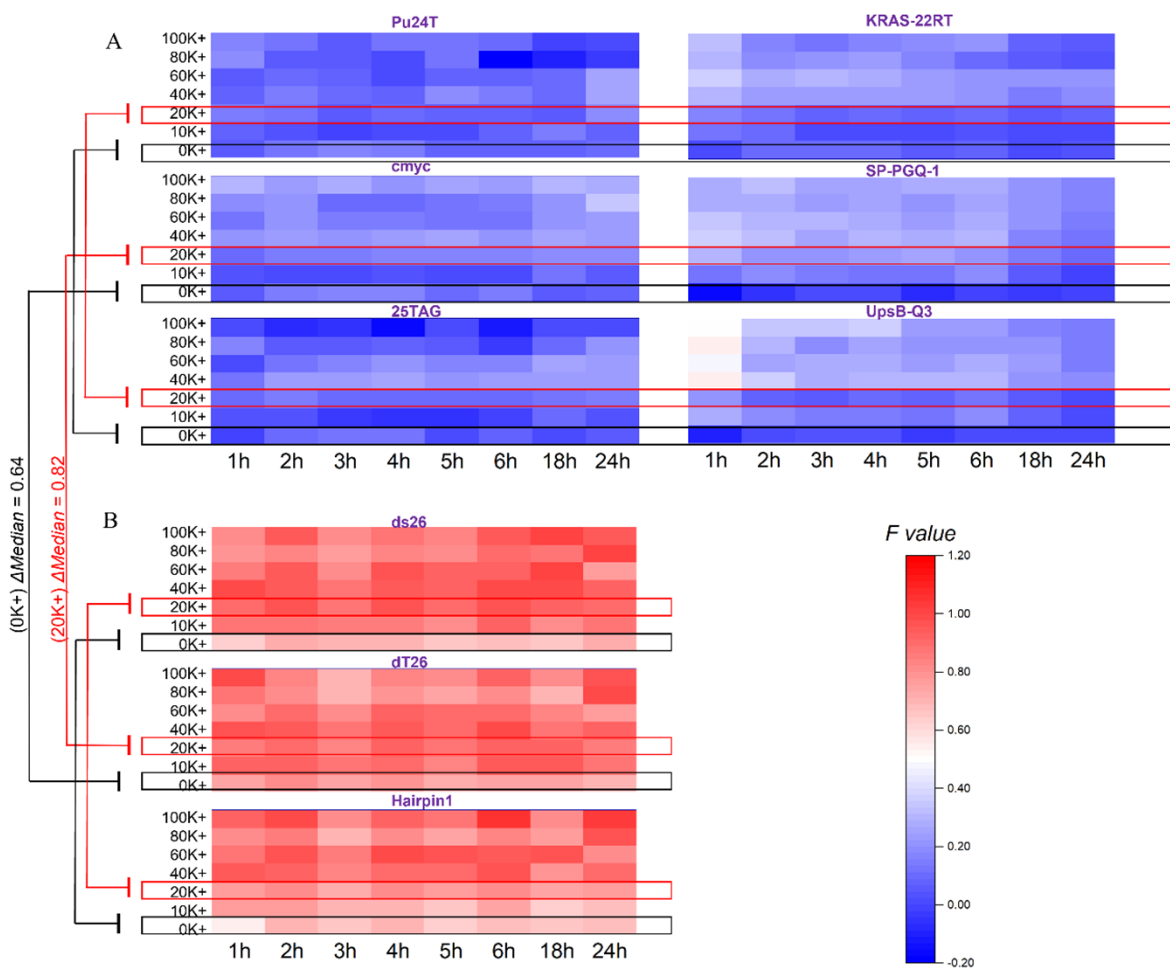


Figure S5. Effects of competitors in group A (G4 competitors) and group B (non-G4 competitors) on fluorescence intensity, tested at different times and potassium concentrations. Each well contained 5 μ M competitor, 200 nM 37Q, 1 μ M PhenDC3 and 20 nM F22; control wells (F22 and F22-37Q) included 20 nM F22 in the presence or absence of 200 nM 37Q. All samples were tested in triplicate at various potassium concentrations (0-100 mM) at RT. Both sub-panels use the same *F* value scale.

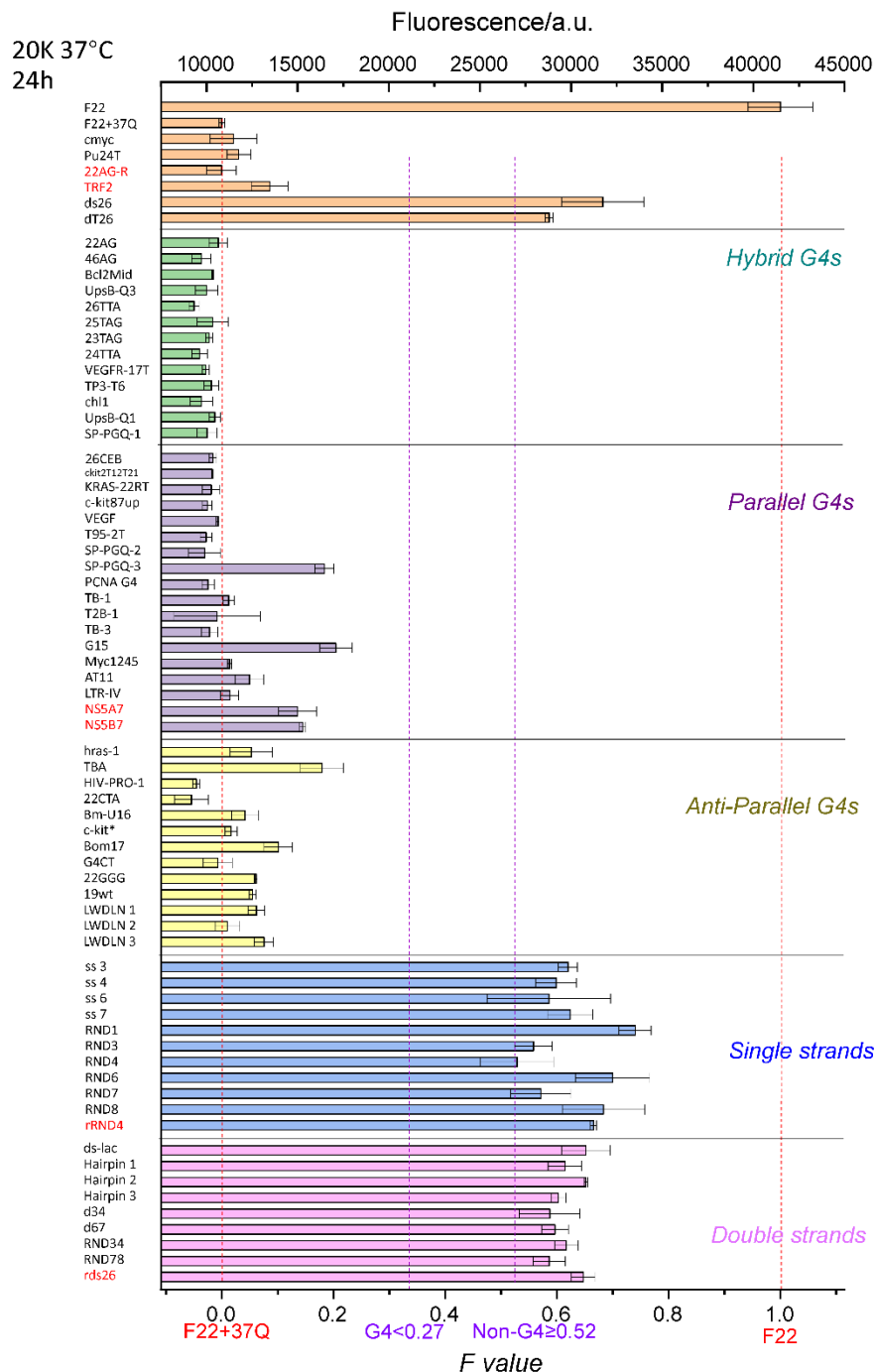


Figure S6. Fluorescence quenching in the presence of various competitors. 20 nM F22 is incubated for 24 h in the presence of 200 nM 37Q and 1 μ M PhenDC3, alone or in the presence of 5 μ M of a variety of competitors, listed on the left (RNAs are marked in red). The F values (bottom X-axis) provides a normalized value. The four different vertical dotted lines correspond to *i*) the level of fluorescence in the absence of a competitor ($F = 0$); *ii*) the first threshold value at 0.27 chosen for positive samples (G4-forming sequences all exhibit F values between 0 and 0.27); *iii*) the second threshold value at 0.52: negative controls (non G4-forming sequences) all exhibit F values between 0.52 and 1; *iv*) the level of fluorescence of F22 alone, with no 37Q added ($F = 1$). Samples were measured in 20K buffer at 37 °C.

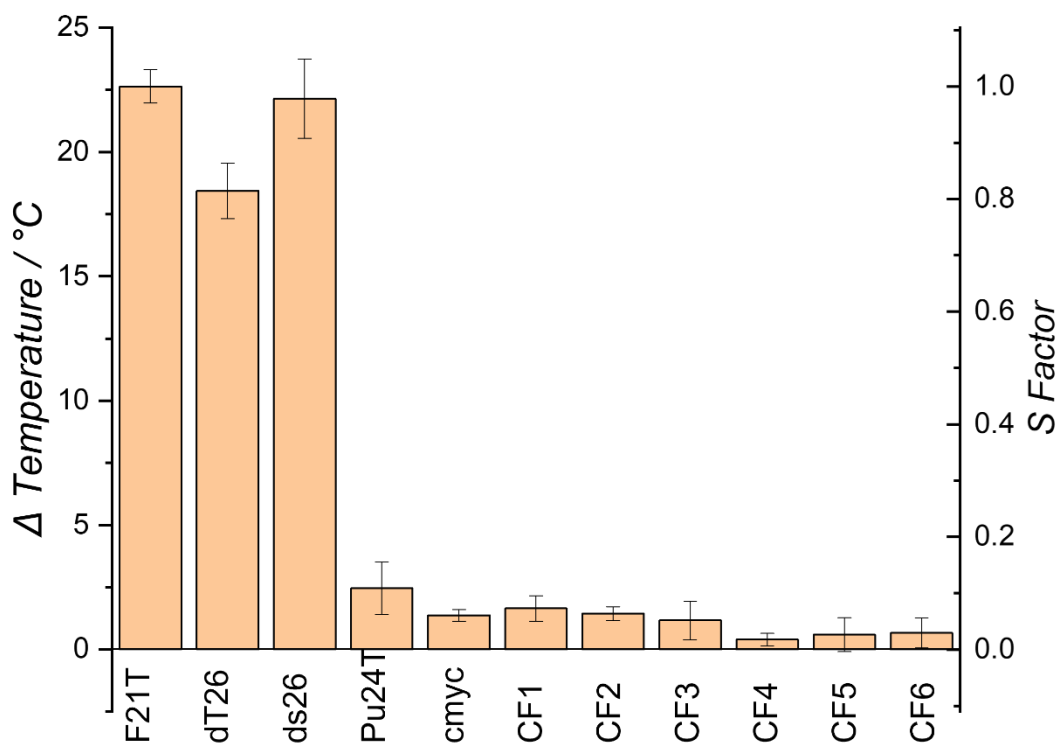


Figure S7. ΔT_m induced by 0.4 μM PhenDC3 on 0.2 μM F21T, alone or in the presence of 3 μM CF series competitors. The *S Factor* (right Y-axis) provides a normalized value. Samples were annealed and measured in 10K buffer (10 mM KCl, 90 mM LiCl, 10 mM lithium cacodylate, pH 7.2).

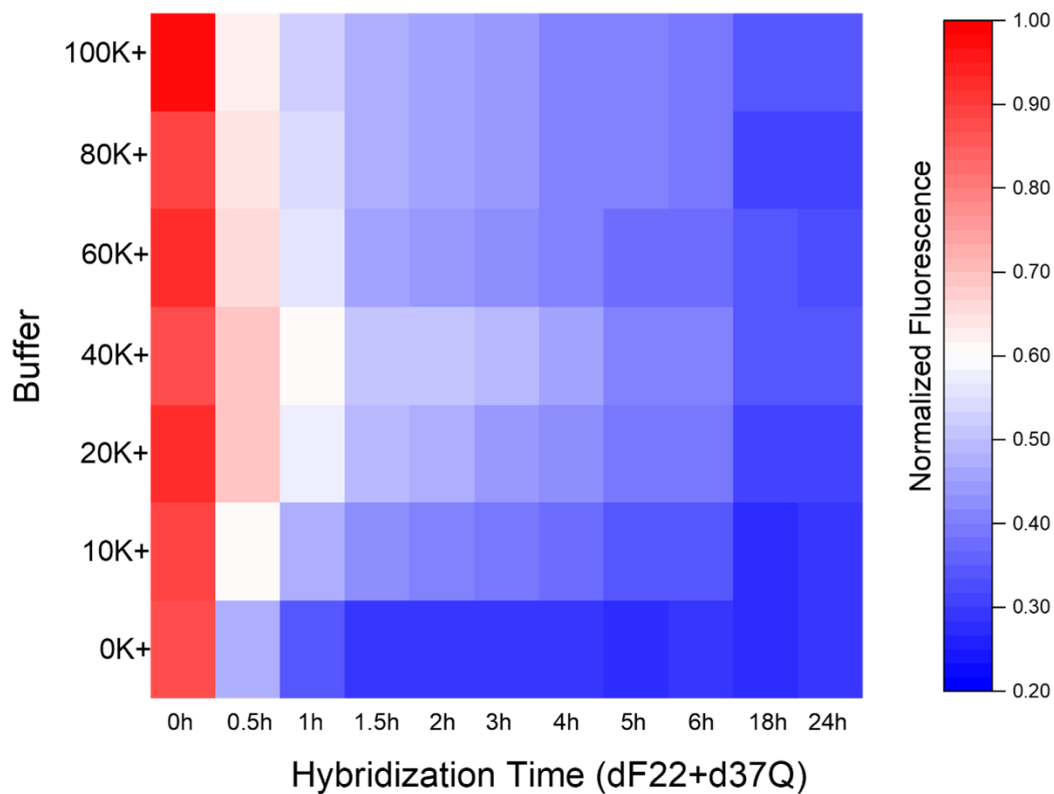


Figure S8. Time-dependent fluorescence of the system involving fluorescently-labeled DNA oligonucleotides (dF22+d37Q) at different potassium concentrations. Concentration of dF22 and d37Q were 20 nM and 200 nM, respectively. d37Q was kept in corresponding buffer for 5 min before adding dF22. Hybridizations were processed at RT.

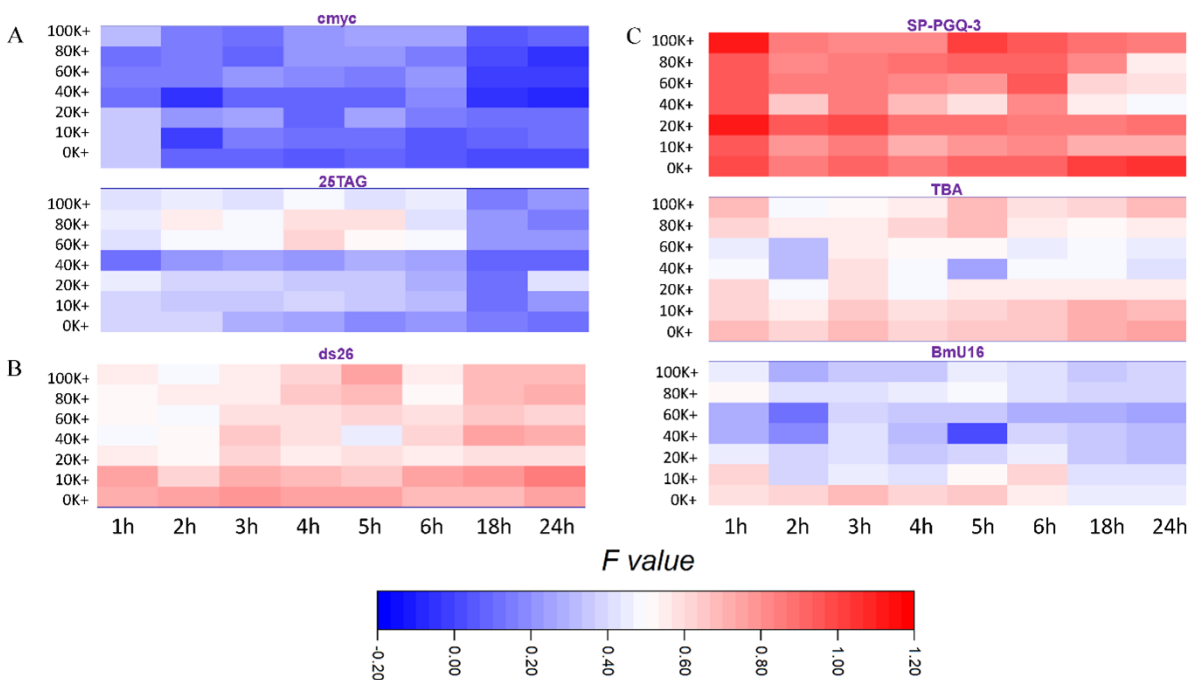


Figure S9. Effects of competitors in group A (G4 competitors), group B (non-G4) and group C (mediocre G4 competitors) on fluorescence intensity, tested at different times and potassium concentrations. Each well contained 5 μ M competitor, 200 nM d37Q, 1 μ M PhenDC3 and 20 nM dF22; control wells (dF22 and dF22-d37Q) included 20 nM dF22 in the presence or absence of 200 nM d37Q. All samples were tested in triplicate at various potassium concentrations (0-100 mM) at RT. All sub-panels are related to the same scale.

A

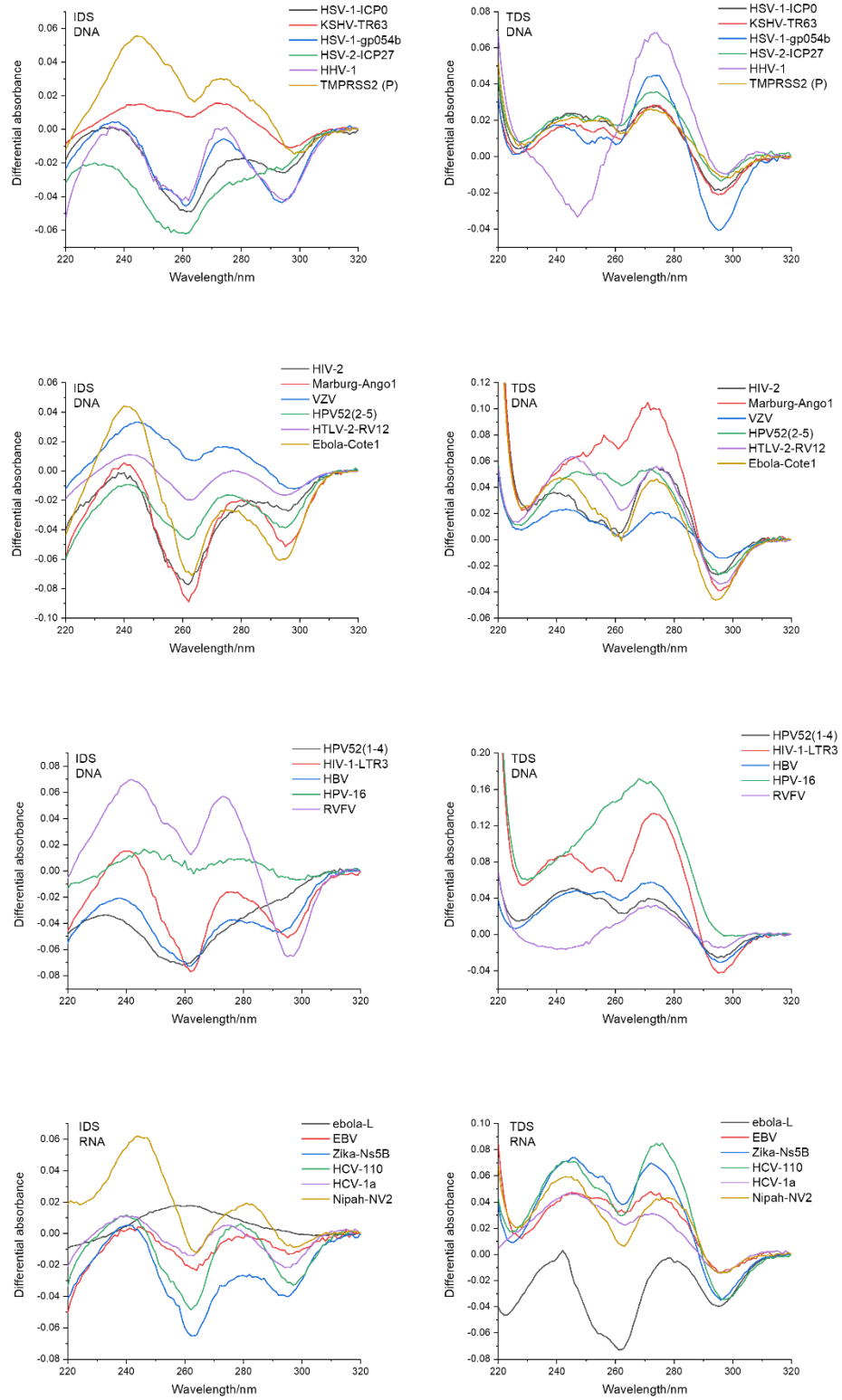


Figure S10. To be continued.

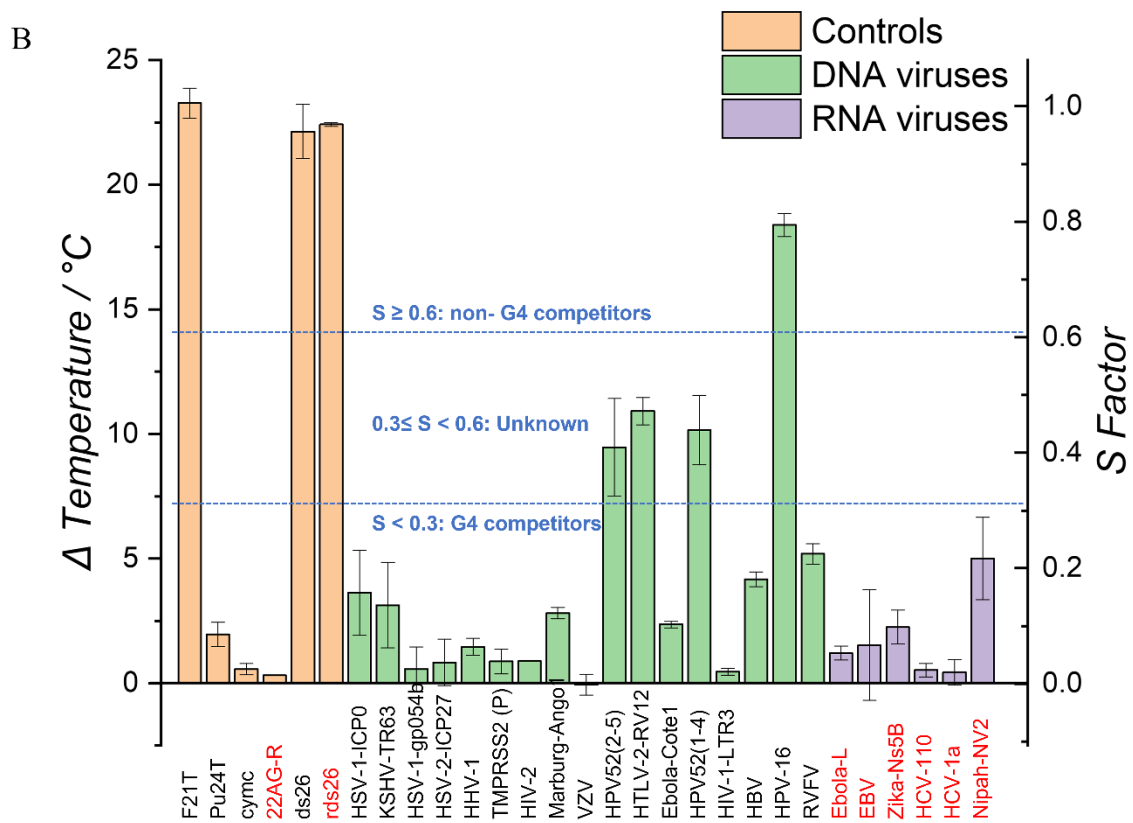


Figure S10. To be continued.

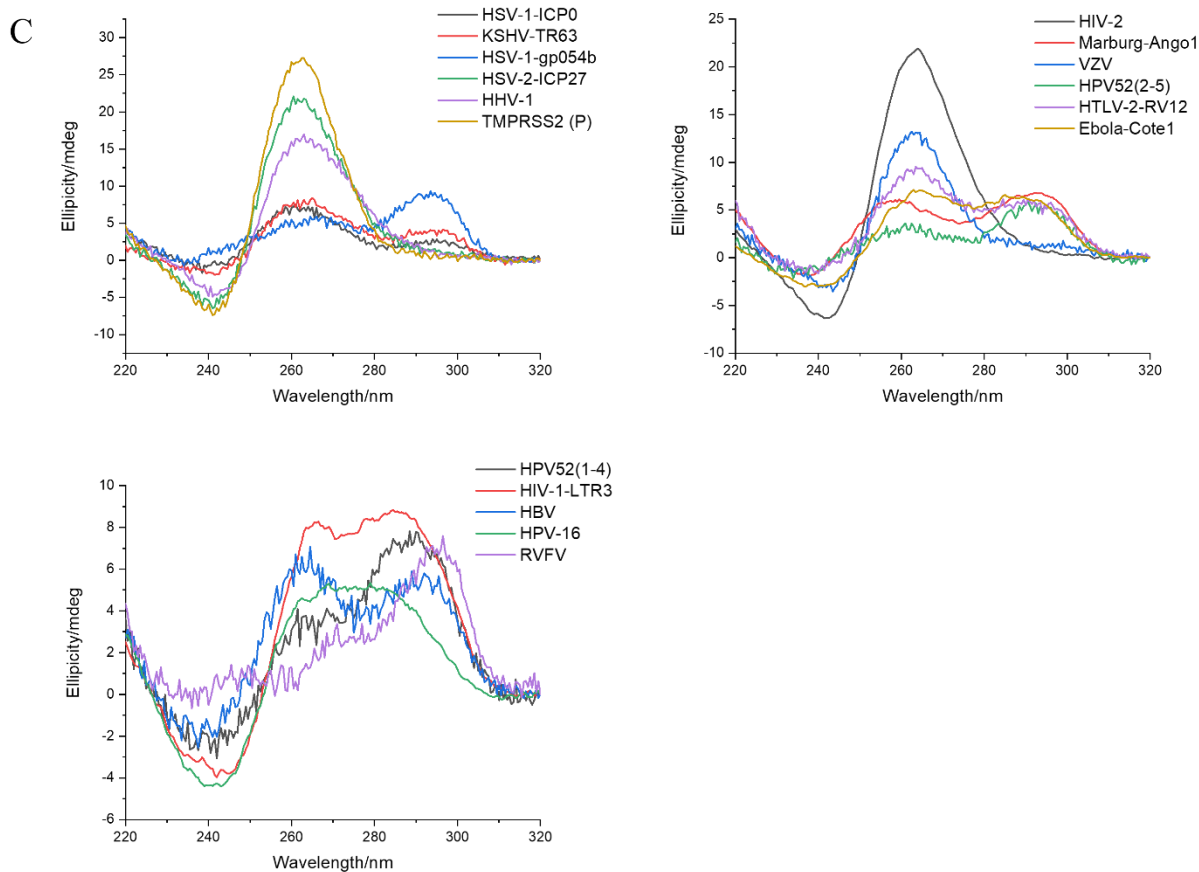


Figure S10. (A) Differential absorbance spectra of 3 μM pre-folded G4-prone sequences: IDS (Isothermal differential absorbance, left spectra) was measured in 10 mM lithium cacodylate pH 7.2 buffer in the absence or presence of 100 mM KCl; TDS (thermal differential absorbance, right spectra) was measured in 100K buffer (100 mM KCl, 10 mM lithium cacodylate, pH 7.2) at 25 $^{\circ}\text{C}$ and 95 $^{\circ}\text{C}$; (B) ΔT_m induced by 0.4 μM PhenDC3 on 0.2 μM F21T, alone or in the presence of 3 μM virus G4-prone competitors. RNA samples are shown in red. The *S Factor* (right Y-axis) provides a normalized value. Samples were annealed and measured in 10K buffer. (C) Circular dichroism spectra of 3 μM pre-folded DNA G4-prone sequences in 100K buffer.

Additional References

1. Stefan, L., F. Denat, and D. Monchaud, *Deciphering the DNAzyme activity of multimeric quadruplexes: insights into their actual role in the telomerase activity evaluation assay*. J Am Chem Soc, 2011. **133**(50): p. 20405-15.
2. Smargiasso, N., V. Gabelica, C. Damblon, F. Rosu, E. De Pauw, M.P. Teulade-Fichou, J.A. Rowe, and A. Claessens, *Putative DNA G-quadruplex formation within the promoters of Plasmodium falciparum var genes*. BMC Genomics, 2009. **10**: p. 362.
3. Mishra, S.K., N. Jain, U. Shankar, A. Tawani, T.K. Sharma, and A. Kumar, *Characterization of highly conserved G-quadruplex motifs as potential drug targets in Streptococcus pneumoniae*. Sci Rep, 2019. **9**(1): p. 1791.
4. Redstone, S.C.J., A.M. Fleming, and C.J. Burrows, *Oxidative Modification of the Potential G-Quadruplex Sequence in the PCNA Gene Promoter Can Turn on Transcription*. Chem Res Toxicol, 2019. **32**(3): p. 437-446.
5. Mukundan, V.T. and A.T. Phan, *Bulges in G-quadruplexes: broadening the definition of G-quadruplex-forming sequences*. J Am Chem Soc, 2013. **135**(13): p. 5017-28.
6. Membrino, A., S. Cogoi, E.B. Pedersen, and L.E. Xodo, *G4-DNA formation in the HRAS promoter and rational design of decoy oligonucleotides for cancer therapy*. PLoS One, 2011. **6**(9): p. e24421.
7. Amrane, S., A. Kerkour, A. Bedrat, B. Vialet, M.L. Andreola, and J.L. Mergny, *Topology of a DNA G-quadruplex structure formed in the HIV-1 promoter: a potential target for anti-HIV drug development*. J Am Chem Soc, 2014. **136**(14): p. 5249-52.
8. Lim, K.W., P. Alberti, A. Guedin, L. Lacroix, J.F. Riou, N.J. Royle, J.L. Mergny, and A.T. Phan, *Sequence variant (CTAGGG)_n in the human telomere favors a G-quadruplex structure containing a G.C.G.C tetrad*. Nucleic Acids Res, 2009. **37**(18): p. 6239-48.
9. Amrane, S., R.W. Ang, Z.M. Tan, C. Li, J.K. Lim, J.M. Lim, K.W. Lim, and A.T. Phan, *A novel chair-type G-quadruplex formed by a Bombyx mori telomeric sequence*. Nucleic Acids Res, 2009. **37**(3): p. 931-8.
10. Dudek, M., M. Deiana, K. Szkaradek, M.J. Janicki, Z. Pokladek, R.W. Gora, and K. Matczyszyn, *Light-Induced Modulation of Chiral Functions in G-Quadruplex-Photochrome Systems*. J Phys Chem Lett, 2021. **12**(39): p. 9436-9441.
11. Rehm, C., I.T. Holder, A. Groß, F. Wojciechowski, M. Urban, M. Sinn, M. Drescher, and J.S. Hartig, *A bacterial DNA quadruplex with exceptional K⁺ selectivity and unique structural polymorphism*. Chem. Sci., 2014. **5**(7): p. 2809-2818.
12. Sacca, B., L. Lacroix, and J.L. Mergny, *The effect of chemical modifications on the thermal stability of different G-quadruplex-forming oligonucleotides*. Nucleic Acids Res, 2005. **33**(4): p. 1182-92.
13. Gomez, D., A. Guedin, J.L. Mergny, B. Salles, J.F. Riou, M.P. Teulade-Fichou, and P. Calsou, *A G-quadruplex structure within the 5'-UTR of TRF2 mRNA represses translation in human cells*. Nucleic Acids Res, 2010. **38**(20): p. 7187-98.
14. Zhang, X., Y. Wei, T. Bing, X. Liu, N. Zhang, J. Wang, J. He, B. Jin, and D. Shangguan, *Development of squaraine based G-quadruplex ligands using click chemistry*. Sci Rep, 2017. **7**(1): p. 4766.
15. Frasson, I., M. Nadai, and S.N. Richter, *Conserved G-Quadruplexes Regulate the Immediate Early Promoters of Human Alpha herpesviruses*. Molecules, 2019. **24**(13).
16. Madireddy, A., P. Purushothaman, C.P. Loosbroock, E.S. Robertson, C.L. Schildkraut, and S.C. Verma, *G-quadruplex-interacting compounds alter latent DNA replication and episomal persistence of KSHV*. Nucleic Acids Res, 2016. **44**(8): p. 3675-94.
17. Artusi, S., M. Nadai, R. Perrone, M.A. Biasolo, G. Palu, L. Flamand, A. Calistri, and S.N. Richter, *The Herpes Simplex Virus-1 genome contains multiple clusters of repeated G-quadruplex: Implications for the antiviral activity of a G-quadruplex ligand*. Antiviral Res, 2015. **118**: p. 123-31.
18. Biswas, B., P. Kumari, and P. Vivekanandan, *Pac1 Signals of Human Herpesviruses Contain a Highly Conserved G-Quadruplex Motif*. ACS Infect Dis, 2018. **4**(5): p. 744-751.

19. Shen, L.W., M.Q. Qian, K. Yu, S. Narva, F. Yu, Y.L. Wu, and W. Zhang, *Inhibition of Influenza A virus propagation by benzoselenoxanthenes stabilizing TMPRSS2 Gene G-quadruplex and hence down-regulating TMPRSS2 expression*. *Sci Rep*, 2020. **10**(1): p. 7635.
20. Krafcikova, P., E. Demkovicova, A. Halaganova, and V. Viglasky, *Putative HIV and SIV G-Quadruplex Sequences in Coding and Noncoding Regions Can Form G-Quadruplexes*. *J Nucleic Acids*, 2017. **2017**: p. 6513720.
21. Krafcikova, P., E. Demkovicova, and V. Viglasky, *Ebola virus derived G-quadruplexes: Thiazole orange interaction*. *Biochim Biophys Acta Gen Subj*, 2017. **1861**(5 Pt B): p. 1321-1328.
22. Tluckova, K., M. Marusic, P. Tothova, L. Bauer, P. Sket, J. Plavec, and V. Viglasky, *Human papillomavirus G-quadruplexes*. *Biochemistry*, 2013. **52**(41): p. 7207-16.
23. Ruggiero, E., M. Tassinari, R. Perrone, M. Nadai, and S.N. Richter, *Stable and Conserved G-Quadruplexes in the Long Terminal Repeat Promoter of Retroviruses*. *ACS Infect Dis*, 2019. **5**(7): p. 1150-1159.
24. Marusic, M. and J. Plavec, *Towards Understanding of Polymorphism of the G-rich Region of Human Papillomavirus Type 52*. *Molecules*, 2019. **24**(7).
25. Perrone, R., M. Nadai, I. Frasson, J.A. Poe, E. Butovskaya, T.E. Smithgall, M. Palumbo, G. Palu, and S.N. Richter, *A dynamic G-quadruplex region regulates the HIV-1 long terminal repeat promoter*. *J Med Chem*, 2013. **56**(16): p. 6521-30.
26. Biswas, B., M. Kandpal, and P. Vivekanandan, *A G-quadruplex motif in an envelope gene promoter regulates transcription and virion secretion in HBV genotype B*. *Nucleic Acids Res*, 2017. **45**(19): p. 11268-11280.
27. Marusic, M., L. Hosnjak, P. Krafcikova, M. Poljak, V. Viglasky, and J. Plavec, *The effect of single nucleotide polymorphisms in G-rich regions of high-risk human papillomaviruses on structural diversity of DNA*. *Biochim Biophys Acta Gen Subj*, 2017. **1861**(5 Pt B): p. 1229-1236.
28. Charley, P.A., C.J. Wilusz, and J. Wilusz, *Identification of phlebovirus and arenavirus RNA sequences that stall and repress the exoribonuclease XRN1*. *J Biol Chem*, 2018. **293**(1): p. 285-295.
29. Wang, S.R., Q.Y. Zhang, J.Q. Wang, X.Y. Ge, Y.Y. Song, Y.F. Wang, X.D. Li, B.S. Fu, G.H. Xu, B. Shu, P. Gong, B. Zhang, T. Tian, and X. Zhou, *Chemical Targeting of a G-Quadruplex RNA in the Ebola Virus L Gene*. *Cell Chem Biol*, 2016. **23**(9): p. 1113-1122.
30. Murat, P., J. Zhong, L. Lekieffre, N.P. Cowieson, J.L. Clancy, T. Preiss, S. Balasubramanian, R. Khanna, and J. Tellam, *G-quadruplexes regulate Epstein-Barr virus-encoded nuclear antigen 1 mRNA translation*. *Nat Chem Biol*, 2014. **10**(5): p. 358-64.
31. Fleming, A.M., Y. Ding, A. Alenko, and C.J. Burrows, *Zika Virus Genomic RNA Possesses Conserved G-Quadruplexes Characteristic of the Flaviviridae Family*. *ACS Infect Dis*, 2016. **2**(10): p. 674-681.
32. Jaubert, C., A. Bedrat, L. Bartolucci, C. Di Primo, M. Ventura, J.L. Mergny, S. Amrane, and M.L. Andreola, *RNA synthesis is modulated by G-quadruplex formation in Hepatitis C virus negative RNA strand*. *Sci Rep*, 2018. **8**(1): p. 8120.
33. Wang, S.-R., Y.-Q. Min, J.-Q. Wang, C.-X. Liu, B.-S. Fu, F. Wu, L.-Y. Wu, Z.-X. Qiao, Y.-Y. Song, G.-H. Xu, Z.-G. Wu, G. Huang, N.-F. Peng, R. Huang, W.-X. Mao, S. Peng, Y.-Q. Chen, Y. Zhu, T. Tian, X.-L. Zhang, and X. Zhou, *A highly conserved G-rich consensus sequence in hepatitis C virus core gene represents a new anti-hepatitis C target*. *Science Advances*. **2**(4): p. e1501535.
34. Majee, P., S. Kumar Mishra, N. Pandya, U. Shankar, S. Pasadi, K. Muniyappa, D. Nayak, and A. Kumar, *Identification and characterization of two conserved G-quadruplex forming motifs in the Nipah virus genome and their interaction with G-quadruplex specific ligands*. *Sci Rep*, 2020. **10**(1): p. 1477.

Chapter IV. G-quadruplexes characterization *in vitro*

We presented in the introduction several methods to characterize G-quadruplexes, and the previous two chapters introduced two novel FRET-based high throughput assays. Based on these considerations, we can now provide a general comparison of G4 characterization methods *in vitro* (**Table 6**). Since all methods have been repeatedly validated, and work well in most cases. we can provide general guidelines for G4 characterization, with an emphasis on the limitations of each method to exclude examples which are not suitable, and provide minimum time- and samples-consuming protocols.

To exhibit detail experiment processes and G4 specificities in different biophysical characterizations, we prepared a guideline paper on how to study G4 structures *in vitro*:

Review for G4 characterization *in vitro* *Guidelines for G-quadruplexes: I. in vitro characterization.* **Yu Luo**, Anton Granzhan, Julien Marquevielle, Anne Cucchiarini, Laurent Lacroix, Samir Amrane, Daniela Verga, Jean-Louis Mergny. (*Biochimie, Under revision*)

Table 6 Comparison of biophysical G4 characterizations in vitro

Methods	G4 specificity	Experiment time*	Samples (pmol)	Limitations
Isothermal absorbance spectra (IDS)	Negative peak at 295 nm	~10 min	3,000	Strong G4s may fold without ions, leading false negative
Thermal absorbance spectra (TDS)	Negative peak at 295 nm	~10 min	3,000	G4s with very high T_m may not be followed by TDS; other structures may also present negative peaks around 295 nm on TDS (see Table 3)
UV melting	Reversed sigmoid melting curve at 295 nm	~ 14 hours	3,000	Absorbance signal at 295 nm may be disturbed by other structures
CD spectra	Specific peaks depending on topology (Table 4)	~15 min	3,000	Parallel G4 CD pattern may be confounded by other structures (see Table 4)
1D ^1H NMR spectra	Peaks in 10 - 12 ppm	≥ 2 hours	350,000	Hoogsteen bps such as GG or GU also present peaks in the range of 10 – 12 ppm.
Gel electrophoresis	Migration speed; G4 ligands specific stain	≥ 4 hours	150	Gels easily get hot in long time migrations
Intrinsic fluorescence	300 - 400 nm	~ 15 min	5,000	High-quality spectrofluorometer (S/N $\geq 2000:1$) required
Fluorescence 'light-up' assay	Fluorescence increase	~ 5 min (48 samples)	75	No perfect fluorescent G4 ligands so far (see Table 5)
FRET-MC	$S < 0.3$	≤ 2 hours (48 samples)	75	Caution with sequences have T_m lower than F21T
Iso-FRET	$F < 0.33$ at 25 °C	thousands of samples per day	125	Caution with sequences have high complementarity level to F22

* Experiment time exclude the samples preparation.

Chapter V. A sodium / potassium switch for G4-prone GC-rich sequences

Metal ions are essential components for sustaining living creatures. The most prevalent metal ions in the human body - as well as in most living species - are potassium and sodium. Sodium is one of the most crucial electrolytes in the extracellular fluid, while potassium is mainly an intracellular ion. The imbalance of sodium and potassium has been related to several diseases. As G4 are ion-dependent structures, a change in $[Na^+] / [K^+]$ ratio may affect genomic G4 folding. Some studies suggested that the unfolding of G4s in $[K^+]$ deficient environment correlated to oncogene expression [245] and pre-mRNA alternative splicing [246]. However, most *in vitro* studies focused on characterizing G4 formation and stability are performed in monocationic buffers (potassium or sodium), which are far from the real intracellular conditions.

Cytosines can pair to guanines to form stable GC base pairs and disrupt G4 structures, and our bioinformatic results showed that about 10% of PQS in the human genome (hg19) contain at least one run of three continuous cytosines. This observation triggered our interest to study how $[Na^+] / [K^+]$ ratio influence the structures of GC-rich sequences. We chose a well-known human minisatellite G4 structure, CEB25wt, as a model sequence. CEB25wt has 9 nt long central loop (CEB25, PDB: 2LPW), allowing us to introduce multiple bases substitutions in this region. We replaced CEB25wt loop bases with cytosines, with the number of three-continuous-cytosine motifs (N) increased from 1 to 4. The total ionic strength was kept constant, with a total concentration of $[Na^+]$ and $[K^+]$ of 140 mM. We set nine buffers with different ratios of $[Na^+] / [K^+]$. UV-melting and CD spectra have been used to characterize the predominant structures of these GC-rich sequences in different $[Na^+] / [K^+]$ buffers. Our results evidenced that the presence of cytosines in G4 loops does not prevent G4 folding or decrease G4s stability, but increases the probability of forming a competing structure, either a hairpin or a duplex. The T_m values of these G4 structures only depend on the specific potassium / sodium ratios, rather than on the specific cytosine-content of the mutated sequences. When $N \leq 2$, G4 structure is the primary structure when potassium ion concentration is sufficient (40 - 140 mM). NMR spectra and DNAzyme activity probe determined that the structures of CEBm4 ($N = 2$) are closely related to $[Na^+] / [K^+]$ ratios: it mainly folded into a hairpin in pure $[Na^+]$ buffer and in a quadruplex in $[K^+]$ buffer. When $N \geq 3$, CEB mutated sequence predominately folds into a duplex under all ionic conditions, and T_m values of duplexes are barely affected by changes in $[K^+] / [Na^+]$ ratios and only depend on the primary sequence (GC content).

A sodium / potassium switch for G4-prone GC-rich sequences

Manuscript in preparation; December 18 version.

Yu Luo^{1,4}, Martina Lenarčič Živković^{2,3} (...)
Daniela Verga^{4,5*}, Lukáš Trantírek^{2*} & Jean-Louis Mergny^{1*}

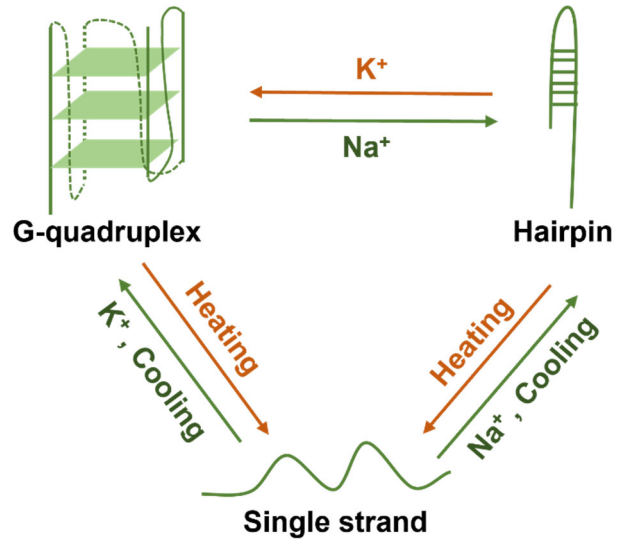
1. Laboratoire d'Optique et Biosciences, Ecole Polytechnique, CNRS, Inserm, Institut Polytechnique de Paris, 91128 Palaiseau, France;
2. Central European Institute of Technology, Masaryk University, 625 00 Brno, Czech Republic;
3. Slovenian NMR Centre, National Institute of SI-1000 Chemistry Ljubljana, Slovenia;
4. CNRS UMR9187, INSERM U1196, Université Paris-Saclay, F-91405 Orsay, France;
5. CNRS UMR9187, INSERM U1196, Institut Curie, PSL Research University, F-91405 Orsay, France;

* **Authors to whom correspondance may be addressed:** Daniela.Verga@curie.fr;
lukas.trantirek@ceitec.muni.cz or jean-louis.mergny@inserm.fr

Abstract

Metal ions are essential components for the survival of living organisms. For most species, intracellular and extracellular ionic conditions differ significantly. As G4 is an ion-dependent structure, changes in the $[\text{Na}^+] / [\text{K}^+]$ ratio may affect the folding of genomic G4. More than 11,000 G4 putative G4 sequences in the human genome (hg19) contain at least two runs of three continuous cytosines, and these mixed G/C rich sequences may form a quadruplex or a competing hairpin structure based on G-C base pairing. In this study, we examine how the $[\text{Na}^+] / [\text{K}^+]$ ratio influences the structures of G C-rich sequences. The natural G4 structure with a long central loop, CEB25wt, was chosen as a model sequence, and the loop bases were gradually replaced by cytosines. The series of CEB mutations revealed that the presence of cytosines in G4 loops does not prevent G4 folding or decrease G4s stability, but increases the probability of forming a competing structure, either a hairpin or an intramolecular duplex. “Shape-shifting” sequences may be used to monitor $[\text{Na}^+] / [\text{K}^+]$ balance.

Keywords: Quadruplex-duplex equilibrium; ionic conditions; structural switch; cation dependency.



TOC. The structure of GC-rich sequences switched by ions and temperature.

1. Introduction

Metal ions are essential components for sustaining plant, animal, and human life. They participate in multiple metabolic processes and are present in all living creatures. Metal ions are involved in intracellular and intercellular communication, the preservation of electrical charges and osmotic pressure, photosynthesis and electron transfer activities and, for nucleic acids, the maintenance of base pairing and stacking [1]. The most prevalent metal ions in the human body - as well as in most living species - are potassium and sodium. Sodium is one of the most crucial electrolytes in the extracellular fluid, while potassium is mainly an intracellular ion. The sodium-potassium ATP pumps (Na^+/K^+ -ATPase), located in the plasma membrane of almost every cell, export three sodium ions in exchange of two potassium ions, and this active transport process is primarily responsible for controlling the balance between sodium and potassium [2]. Potassium / sodium balance helps to maintain vital body functions. In contrast, potassium / sodium imbalance contributes to several diseases. Sodium intake has been tightly related to blood pressure and hypertension [3]. Intracellular $[\text{Na}^+]$ and $[\text{K}^+]$ levels are both increased in brain regions of patients affected by Alzheimer disease [4].

G-quadruplexes (G4s) are four-stranded nucleic acid secondary structures constituted by two or more stacked G-quartets. G4s have been shown to be involved in DNA replication [5], gene transcription [6] and pre-mRNA alternative splicing [7]. As topology and stability of G4 structures are sensitive to the cation type and ionic strength [8], putative quadruplex sequences (PQS) in different potassium / sodium ratio environments may adopt distinct conformations, resulting in diverse effects on cell activities. During tumor progression, G4 structures have been demonstrated to influence gene expression in response to changes in intracellular cation concentrations. Due to the overexpression of a $[\text{K}^+]$ channel, malignant cancer cells (*i.e.*, highly metastatic breast cancer cells MDA-MB-231) may exhibit a significantly lower intracellular $[\text{K}^+]$ than normal cells. In normal cells, high $[\text{K}^+]$ inhibits the transcription of certain oncogenes by stabilizing G-quadruplex structures, whereas in cancer cells lower $[\text{K}^+]$ has been proposed to destabilize G4s and increase the transcription of oncogenes [9]. This result is surprising, given that cancer cells seem to exhibit more G4s, as shown by BG4 antibody immunostaining experiments [10]. A recent study reported that potassium / sodium balance is important to regulate alternative promoter usage and/or pre-mRNA splicing

in the transcription of SGK1, *via* the folding of a G4 structure located in its promoter. The SGK1 gene has a G4-forming motif located upstream to the proximal region of promoter-2, which has been shown to be stabilized by potassium ions and resveratrol but destabilized by sodium ions. For all three SGK1 isoforms, transcription is stimulated by high levels of sodium, whereas resveratrol or potassium ion addition suppresses the transcription of isoform-2 and isoform-3, but not isoform-1 [11].

Characterizing the structures formed by G-rich strands in multi-ion environments may help comprehend how G4 folding influences gene activities. However, usually, G4 formation is studied under simple monocationic conditions, which differ from the real intracellular conditions. In addition, a number of potential quadruplex sequences contain runs of cytosines directly upstream or downstream of the G-rich sequence, or in the loops. For example, we found that about 10% of PQS in the human genome (hg19) comprised at least three continuous cytosines. These mixed G/C rich sequences may form a quadruplex or a competing hairpin structure based on G-C base pairing. The relative stabilities of these competing structures should depend on the sequence, but also on ionic conditions, given that G4 stability should be dependent on potassium / sodium balance while hairpin duplexes should not.

In this study, we worked on variants of the well-known human minisatellite G4 structure CEB25wt in which we progressively inserted runs of consecutive cytosines. The structure of the CEB25wt quadruplex was solved by NMR (PDB: 2LPW): it corresponds to a parallel intramolecular DNA G4 with a 9 nt long central loop [12]. We analyzed G4 and hairpin formation under different potassium / sodium ion ratios for CEB25wt and its mutants. Starting from this sequence, we introduced base substitutions to increase the share of cytosines in CEB25wt variants in order to identify the factors affecting the equilibrium between G4 and hairpin formation under physiological conditions. We identified “shape-shifting” sequences which respond to K^+/Na^+ balance and are expected to adopt different folds in the intra- and inter-cellular environment.

2. Materials and methods

Materials and samples

Oligonucleotides were purchased from Eurogentec (Belgium) and used without further purification; their sequences are shown in

Table 1. Stock solutions were prepared at 100 μ M strand concentration in milli-Q H₂O. All oligonucleotides were annealed in corresponding buffers, kept at 95 °C for 5 min and slowly cooled to room temperature before measurements. All chemical reagents, including Hemin, ABTS²⁻ and H₂O₂, were purchased from Sigma-Aldrich (France).

Table 1 DNA mutations based on CEB25wt^a

Acronym	Sequence (5' - 3')	Length (nt)	G4H score	N ^b
CEBwt	AAGGGTGGGTGTAAGTGTGGGTGGGT	26	1.50	0
CEBm0A	AAGGGTGGGTAAAAAATGGGTGGGT	26	1.38	0
CEBm0T	AAGGGTGGGTTTTATTTGGGTGGGT	26	1.38	0
CEBm1	AAGGGTGGGTCCCAGTGTGGGTGGGT	26	1.12	1
CEBm2	AAGGGTGGGTCCCACCTGGGTGGGT	26	1.04	1.33
CEBm3	AAGGGTGGGTCCCACCGTGGGTGGGT	26	0.92	1.67
CEBm4	AAGGGTGGGTCCCACCCTGGGTGGGT	26	0.69	2
CEBm5	AAGGGTGGGTCCCACCCTGGGTGGGTCCCA	30	0.30	3
CEBm6	AAGGGTGGGTCCCAGTGTGGGTGGGTCCCACCCA	34	0.32	3
CEBm7	AAGGGTGGGTCCCACCCTGGGTGGGTCCCACCCA	34	0	4

^a The cytosines introduced in the mutant sequences are marked in red; other base substitutions are shown in blue.

^b 'N' represents numbers of three-continuous-cytosines ('CCC') motifs.

Size-Exclusion High-Performance Liquid Chromatography (SE-HPLC)

SE-HPLC experiments were performed on a ÄKTA FPLC 900 system (GE Healthcare, France) equipped with a Frac-900 autosampler, a Thermo Acclaim SEC-300 column (4.6 × 300 mm; 5 μ m hydrophilic polymethacrylate resin spherical particles with 300 Å pore size), and a diode array detector. A solution containing 40 mM cacodylate buffer, pH 6, with 100 mM potassium was used as an elution buffer and to dissolve oligonucleotides. 20 μ L of a 50 μ M oligonucleotide solution in 50 mM Tris·HCl (pH = 7.2) with 140 mM NaCl or 140 mM KCl was injected onto the column (0.15 mL/min elution flow rate at C), and elution was monitored by measuring the absorbance at 260 nm.

UV-melting

Oligonucleotides (5 μM strand concentration) were pre-annealed in 10 mM lithium cacodylate buffer (pH = 7.2) complemented with different concentrations of potassium / sodium ions. We defined nine different buffer conditions, from B1 to B9, as follows:

B1: 140 mM Na^+ / 0 mM K^+ ; B2: 135 mM Na^+ / 5 mM K^+ ; B3: 125 mM Na^+ / 15 mM K^+ ;
B4: 115 mM Na^+ / 25 mM K^+ ; B5: 100 mM Na^+ / 40 mM K^+ ; B6: 80 mM Na^+ / 60 mM K^+ ;
B7: 70 mM Na^+ / 70 mM K^+ ; B8: 20 mM Na^+ / 120 mM K^+ ; B9: 0 mM Na^+ / 140 mM K^+ .

UV-melting curves were recorded with a Cary 300 (Agilent Technologies, France) spectrophotometer. Heating runs were performed between 10 $^{\circ}\text{C}$ and 95 $^{\circ}\text{C}$, the temperature was increased by 0.2 $^{\circ}\text{C}$ per minute, and absorbance was recorded at 260 and 295 nm. T_m was determined as the temperature corresponding to half of the height of the melting curve.

Thermal difference spectra (TDS)

Absorbance spectra of pre-folded CEBm2 in B3 and CEBm3 in B4 (oligonucleotide concentration was 5 μM in a final volume of 1 mL) were recorded on a Cary 300 (Agilent Technologies, France) spectrophotometer at 25 $^{\circ}\text{C}$ and 95 $^{\circ}\text{C}$ (scan range: 400 - 200 nm; scan rate: 600 nm/min; automatic baseline correction). TDS corresponds to the arithmetic difference between the initial (25 $^{\circ}\text{C}$) and second (95 $^{\circ}\text{C}$) spectra.

Circular dichroism (CD) measurements

CD spectra of pre-folded 5 μM oligonucleotides in 1 mL of different buffers were recorded on a JASCO J-1500 (France) spectropolarimeter (scan range: 340 - 200 nm; scan rate: 100 nm/min; averaging three accumulations) at 25 $^{\circ}\text{C}$.

^1H NMR analysis

Spectra were recorded at 25 $^{\circ}\text{C}$ on a 600 MHz NMR instrument in 10 mM lithium cacodylate pH 7.2 10% D_2O buffer. The CEBm4 sequence was tested at 0.2 mM strand concentration. Detailed experiment procedures will be completed by Martina / Lukas.

DNAzyme activity assay

DNA samples were folded by heating at 95 °C for 5 min in corresponding buffer and left to cool at room temperature for at least two hours. Hemin was then added to reach a final concentration of 6 µM and the solutions were left to stand at room temperature (25 °C) for 30 min. ABTS²⁻ was then added to each sample to a final concentration of 500 mM and basal absorbance was measured. To initiate the oxidation reaction, H₂O₂ was added to a final concentration of 50 mM, followed by quick mixing. 15 min later the absorbance at 420 nm of the oxidized product ABTS[•] was monitored by a TECAN M1000 pro plate reader (France). The final sample contained 3 µM DNA, 6 µM Hemin, 1% DMSO, 500 mM ABTS²⁻ and 50 mM H₂O₂ in a total volume of 50 µL.

Methyl mesoporphyrin IX (NMM) light-up probe

In each microwell containing 95 µL of 3.15 µM pre-folded oligonucleotides solubilized in corresponding buffers, 5 µL of NMM at 40 µM were added to reach a final concentration of 2 µM. The microplate was shaken for 1 min and fluorescence was read immediately at 25 °C. Fluorescence intensity was collected at 610 nm after excitation at 380 nm in a TECAN M1000 pro plate reader (France).

3. Results

3.1 Analysis of PQS containing runs of consecutive cytosines

A quadruplex is a four-stranded structure constituted by two or more G-tetrads. In this study, we will only consider classical G4 structures potentially involving three tetrads, in line with the original G4-predicting algorithms [13], with the basic requirement of 4 runs of 3 or more continuous guanines. Two-tetrads quadruplexes are often of low thermal stability and will not be considered here [14].

In this study, we assumed that the presence of at least one run of ‘CCC’ motif ($N \geq 1$) may allow the formation of a 3 consecutive GC bps with a neighboring G4 track, possibly interfering with G4 folding. To search for these motifs, we did not use G4Hunter, as it tends to exclude sequences having blocs of consecutive C [15-17] which contribute negatively to the G4Hunter score [18]. As a consequence, the majority of PQS containing C-runs are disfavored by G4Hunter. We therefore used the GSE133379

database, which is a comprehensive listing of PQS motifs in the human genome (hg19) [19], and includes four validated types of G4s: **i)** the most classical G4s: three or more G-tetrads with short loops (1-7 nt) [13]; **ii)** G4s with long loops (8-15 nt) [20]; **iii)** G-stems with bulges [21]; and **iv)** Guanine-vacancy-bearing G4s [22]. Compared to the first human PQS dataset reported in 2005 [13], many more PQS (1,506,353) have been included in GSE133379 (**Table S1**), and the average PQS frequency is 0.48 per 1000 bps in the whole genome. It is worth noting that about 10% of these PQS contain at least one run of three-continuous-cytosine, and more than half of them are located in functional regions of the genome. Particularly, 44.2% of ‘CCC’-containing ($N \geq 1$) PQS are located in introns and exons, implying the significance to study the structure of C-rich PQS in RNA (**Figure S1**). Then, we further increased the stringency of our PQS search to $N \geq 2$, and found over 11,000 PQS which meet this requirement.

3.2 Primary structures of G C-rich sequences in different Na^+ / K^+ buffers

Having established that a number of PQS motifs containing one more runs of cytosines are present in the genome, we wanted to investigate how they could affect G4 formation. Several cytosine-containing G4 structures have already been characterized (*e.g.*, *VEGF* d[CG₄CG₃CCTTG₃CG₄T], PDB: 2M27), implying that the presence of a few cytosines does not affect G4 formation and stability. We therefore considered motifs having longer runs (3 or more) of cytosines in the loops. To do so, we started with a well-known human minisatellite CEB25wt G4 structure involving a long (9 nucleotides) central loop (PDB: 2LPW), allowing us to introduce multiple bases substitutions in this region. With this in mind, we mutated the sequence by replacing loop bases with cytosines; the first mutation contains one run of the ‘CCC’ motif. SE-HPLC was first used to confirm the molecularity of CEB25wt and each mutation. Strong single peaks corresponding to a retention time of around 18 min were recorded, in agreement with the formation of intramolecular structures in potassium and sodium buffers (**Figure S2**). UV-melting profiles and CD spectra were used to determine structure formation and T_m values of CEB25 (wt) and its mutated variants.

As shown in **Figure 1a**, CEBwt ($N = 0$) mainly formed a G4 structure in all buffers, and T_m of G4s increased with potassium concentration. CEBm2 ($N = 1.33$) predominately formed hairpins in low potassium concentration buffers (B1 and B2), while a G4 specific signature in the melting curves could be observed

in buffers containing 25 mM (B3) or higher potassium concentrations (**Figure 1b**). Since the melting profiles of CEBm2 in B3 (as well as CEBm3 in B4) were ambiguous, TDS were recorded to track the G4 formation of CEBm2 and m3 in these buffers. As shown in **Figure S3**, the negative peak at 295 nm evidenced the formation of G4 structures [23]. For the sequences with a high number of cytosines (N = 4 for CEBm7) duplexes predominated in all buffers with very similar melting temperatures (**Figure 1c**). The UV melting profiles of other sequences are shown in **Figure S4**, and the predominant conformation and T_m are gathered in **Table 2**.

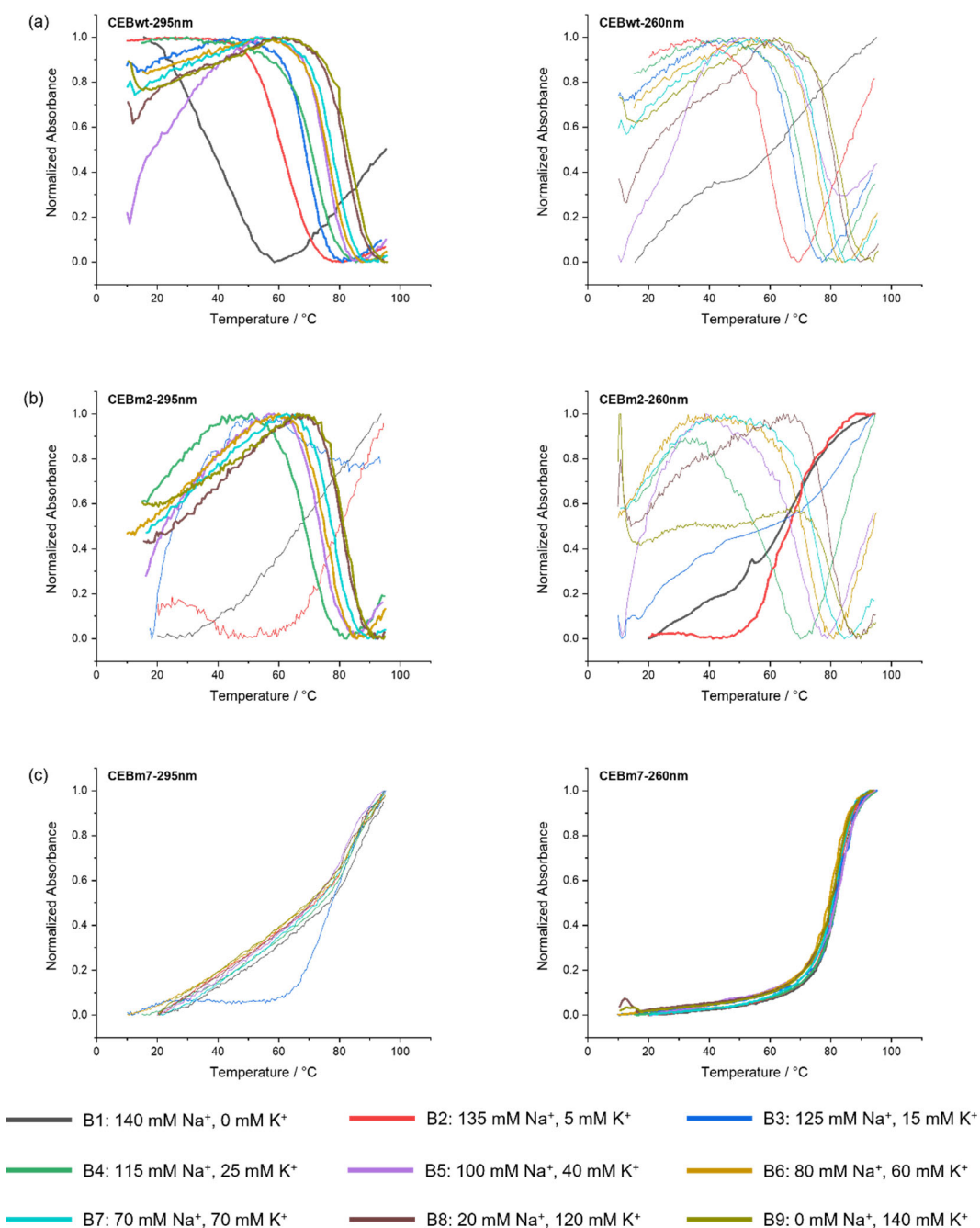


Figure 1. Normalized UV-melting curves of 5 μ M (a) CEBwt, (b) CEBm2 and (c) CEBm7 in different buffers. Heating runs were performed between 10 $^{\circ}$ C and 95 $^{\circ}$ C; the temperature was increased by 0.2 $^{\circ}$ C per minute, and the absorbance was recorded at 260 and 295 nm. For each condition, the curve shown in bold (either corresponding to absorbance at 295 nm or 260 nm) was used to calculate the T_m .

Table 2 T_m (°C) of CEB series DNA strands in different buffers ^a

Acronym	B1	B2	B3	B4	B5	B6	B7	B8	B9
Na ⁺ (mM)	140	135	125	115	100	80	70	20	0
K ⁺ (mM)	0	5	15	25	40	60	70	120	140
CEBwt	45.2	60.5	68.4	70.5	74.3	75.94	77.9	80.9	82.3
CEBm0A	38.9	57.2	65.3	67.6	71.1	73.2	75.6	78.6	79.1
CEBm0T	44.5	59.5	67.8	69.0	72.8	74.8	77.9	79.8	81.7
CEBm1	45.8	60.5	68.5	70.8	74.1	76.2	77.6	80.9	82.3
CEBm2	<i>64.2</i>	<i>65.1</i>	G4^b	69.1	73.3	75.2	77.2	80.5	81.1
CEBm3	<i>69.5</i>	<i>70.8</i>	<i>70.5</i>	G4^b	73.1	74.7	77.1	80.2	80.9
CEBm4	<i>71.5</i>	<i>71.5</i>	<i>71.4</i>	<i>71.0</i>	73.6	74.1	76.9	80.8	81.4
CEBm5	<i>72.2</i>	<i>72.5</i>	<i>72.7</i>	<i>71.2</i>	<i>72.1</i>	<i>71.8</i>	<i>71.5</i>	<i>70.1</i>	<i>70.1</i>
CEBm6	<i>71.9</i>	<i>72.0</i>	<i>71.9</i>	<i>71.9</i>	<i>71.9</i>	<i>70.9</i>	<i>71.0</i>	<i>69.3</i>	<i>69.3</i>
CEBm7	<i>81.7</i>	<i>81.1</i>	<i>81.1</i>	<i>81.7</i>	<i>80.9</i>	<i>80.3</i>	<i>81.5</i>	<i>81.5</i>	<i>81.5</i>

^a Bold red values correspond to quadruplex T_m values calculated from UV melting profiles at 295 nm, while numbers in purple / italics correspond to hairpin duplex melting, calculated from UV melting profiles recorded at 260 nm.

^b G4 structures were evidenced by the negative peak at 295 nm on the TDS spectra, but a precise T_m could not be determined.

Under all buffer conditions (B1 to B9), the predominant conformation adopted by CEBwt and CEBm1 was found to be a G4 structure. In contrast, the structure of the mutated sequences 2 - 4 depends on the specific buffer employed. In a low potassium concentration buffer (B1 - B4, [K⁺] lower than 40 mM), CEBm1 (N = 1) predominantly adopts a G4 structure, even in a buffer that does not containing potassium (B1), but traces of a hairpin could be evidenced by the absorbance profile at 260 nm (**Figure S4c**). For the sequences with an increased number of cytosines, higher concentrations of potassium are required to observe G4 folding. CEBm4, characterized by N = 2, requires 40 mM potassium to fold into a G4 structure. Mutations 2 - 4 mainly fold into G4s at high potassium concentration (B5 - B9, [K⁺] higher than or equal to 40 mM). Interestingly, under conditions where the quadruplex is the predominant fold (sequences CEB25wt to CEBm4 in buffers B5 to B9), its T_m is nearly independent on the number of cytosines in the loops, and we found no significant difference in T_m among the different mutated sequences. This observation illustrates

that cytosines are not detrimental *per se* to G4 formation and that T_m depends on buffer composition rather than in C content, as expected for a G-quadruplex.

Loop composition is an element contributing to G4 folding. In a 1 nt propeller G4 loop, substitution of a nucleotide with a single adenine reduces G4 melting by 6 - 8 °C; differently, loops composed by cytosine, thymine and guanine showed similar T_m [24, 25]. To confirm that central loop base substitutions did not strongly influence G4 intrinsic stability (besides altering G4-hairpin competition), especially in low potassium concentration conditions, we replaced some of the bases in the loops by adenines (CEBm0A) and thymines (CEBm0T). T_m of CEBm0T is very similar to the one of CEBwt, while CEBm0A shows lower thermal stability as compared to CEBwt and CEBm0T, in agreement with the literature [24, 25]. In the mutated C-rich sequences, we replaced one by one the bases present in the central loop by cytosines, and observed that T_m was not affected by loop base substitution. There is currently little information in the literature to explain how base composition in a long loop influences stability of a G4. The small variation in G4 T_m may be associated to the low fraction of adenines present in the central loop (2 adenines in the 9 nt loop [26]). Our results suggested that cytosine substitutions in the central loop of CEBwt barely influence G4 thermal stability, but increase the proportion of hairpins.

CD spectra were recorded to confirm structures and G4 topologies. CEBwt formed a parallel G4 in all buffers (**Figure 2a**). CEBm0A and CEBm0T also showed unambiguous G4 patterns in UV-melting experiments, and CD spectra evidenced that they form parallel G4 structures (**Figure 2b & c**). In contrast, CD spectra of mutated sequences CEBm1- m4 exhibited specific peaks in low potassium concentration buffers (B1 - B4), implying conformation conversion: CEBm1 in B1 showed a positive peak at 240 nm, which shifted to 242 nm in buffers containing higher $[K^+]$ (**Figure 2d**). CEBm2 and CEBm3 shifted the positive peak from 257 nm to 263 nm and 260 nm to 264 nm, respectively, at higher $[K^+]$ (**Figure 2e & f**). A 2 nm bathochromic shift was also observed for CEBm4 at higher $[K^+]$ buffer. Of note, CEBm3 and CEBm4 exhibited dichroic positive signals from 280 nm to 295 nm in low $[K^+]$ buffer; the intensity of the latter decreases gradually with increasing $[K^+]$, and could not be observed in high $[K^+]$ buffers (**Figure 2f**).

& g). These long wavelength maxima may be associated to the formation of GC bps [27, 28], confirming hairpin-G4 K^+/Na^+ dependent conversion.

Previous studies reported that metal ions may lead to G4 topology change. For example, pw17 d(G₃TAG₃CG₃TTG₃) folds into an antiparallel G4 in a buffer containing 5 to 70 mM [Na⁺]; however, the addition of less than 10 mM K⁺ buffer can switch the topology from antiparallel to parallel [29]. The human telomeric G4 d[AGGG(TTAGGG)₃] folds into an antiparallel G4 in 90 mM Na⁺ (PDB: 143D) and into hybrid or parallel G4 in 50 mM K⁺ (PDB: 1KF1). Loop composition may also lead to topology switching. L2T4 d(G₃TG₃T₄G₃TG₃) forms an antiparallel G4 in 100 mM Na⁺, while L2A4 d(G₃TG₃A₄G₃TG₃) folds into a parallel G4 in the same buffer [30]. In contrast, in this work, CD analysis (**Figure 2**) demonstrates that whenever a quadruplex is formed by any of the wt or mutant sequences, it remains parallel, independently of the potassium / sodium ratio. Therefore, the mutations of the loop sequence did not affect G4 topology, as the 1-nt loops impose a chain-reversal character for two of the three loops, severely restricting topological versatility.

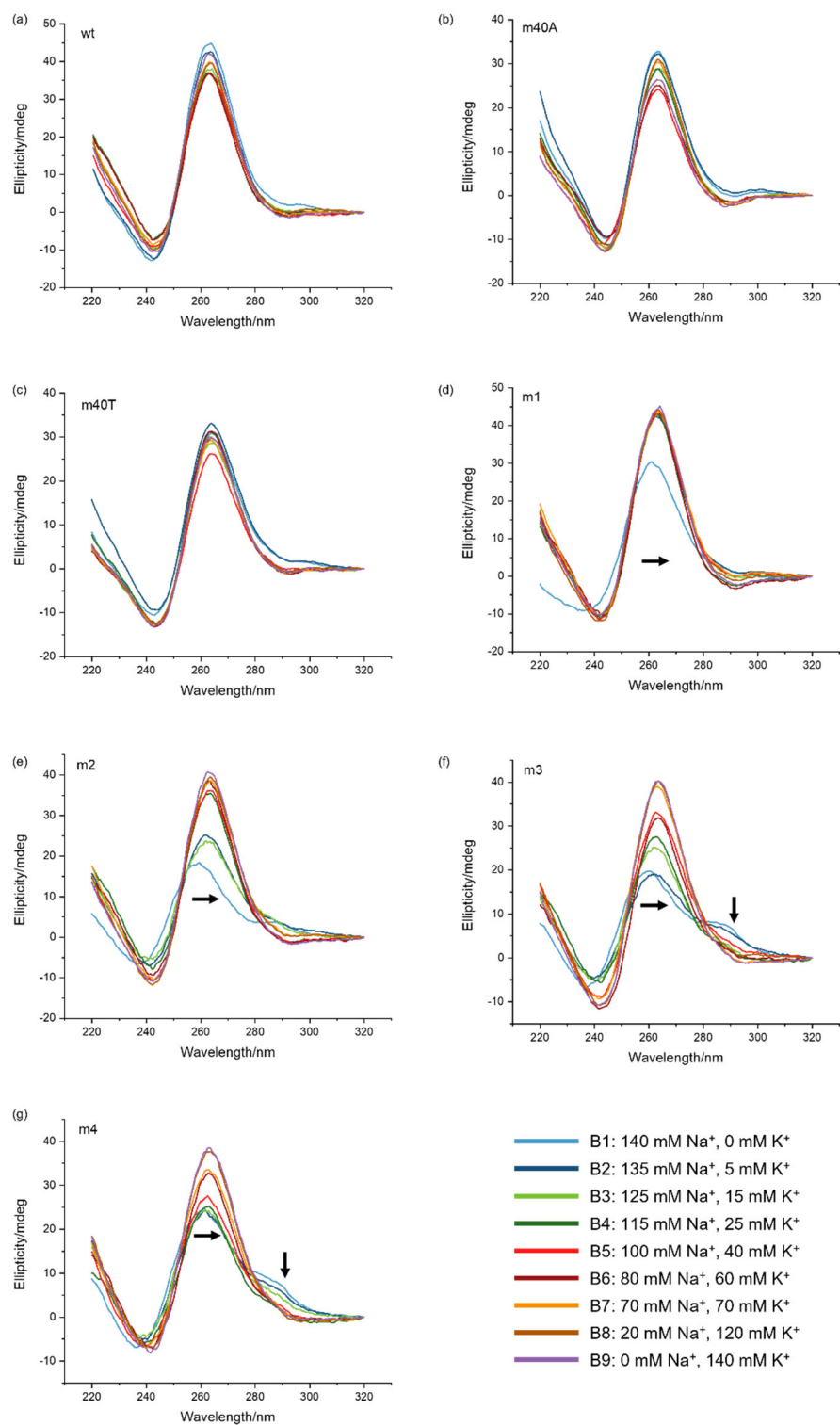


Figure 2. CD spectra of 5 μM CEBwt and its mutated versions in different buffers at 25 $^{\circ}\text{C}$. Arrows indicates peak shifts or peak intensity changes.

When N increased to 3 or higher, CEBm5 - m7 predominately formed duplexes under all ionic conditions. In contrast to G4 structures, their T_m was not affected by changes in $[K^+] / [Na^+]$ ratio, but depended on the expected number of base pairs formed with each specific sequence. Both CEBm5 and m6 allow the formation of up to 9 GC base pairs, but with a different organization of loops and dangling-ends: CEBm5 is expected to adopt a hairpin structure with a relatively small central loop and a 6 mer dangling-end (**Figure S5a**). In contrast, CEBm6 forms a bilaterally symmetrical hairpin with a longer loop (**Figure S5b**). Previous studies have shown that the presence of dangling-ends minimally affect hairpin stability [31], while loop size is an important element affecting hairpin stability. Generally speaking, a 4 nt loop is optimal for hairpin formation in 100 mM $[Na^+]$, and the stability of the hairpin decreases for longer loops [32], as larger loops are expected to have weaker intraloop hydrophobic interactions [32] and stem-loop interactions [33]. In our work, no significant difference in thermal stability was found between CEBm5 and m6, perhaps because of specificities in loop composition [34].

UV-melting results suggest that three 'CCC' motifs are enough to convert G4s to hairpins in all buffers, no matter their specific hairpin conformations. We then replaced the CEBm6 loop bases by cytosines, increasing N to 4. Unsurprisingly, CEBm7 prefers to form a hairpin. Compared to CEBm5 and m6, with a T_m 10 °C higher than CEBm5 and m6, as expected: as G-C bp are more stable than A-T bp.

3.3 CEBm4 as a prototypal example of a Hairpin / Quadruplex switch depending on buffer conditions

UV-melting and CD spectra suggested that the structure of CEBm4 depends on the potassium / sodium ratio: m4 mainly forms a hairpin in low $[K^+]$ conditions and a G4 structure in high $[K^+]$ buffers. To confirm this potassium / sodium dependent structural transition, we performed proton NMR analysis of CEBm4 in three different buffers. **Figure 3** unambiguously confirms that CEBm4 exclusively adopts a hairpin structure in B1 buffer: the imino proton resonances in the region of 12.0 - 14.0 ppm are contributed by the Watson-Crick base pairs [35, 36], implying the formation of hairpin structures. The peaks present in the 10.5 - 12.0 ppm region in 1H NMR spectra are a distinctive feature of G-tetrad formation, therefore, in B9 conditions the resonance of CEBm4 evidenced a highly predominant G4 fold. The imino proton resonances

in B7 showed peaks in both 12.0 - 14.0 ppm and 10.5 - 12.0 ppm regions, suggesting the co-existence of hairpin and G4 structures.

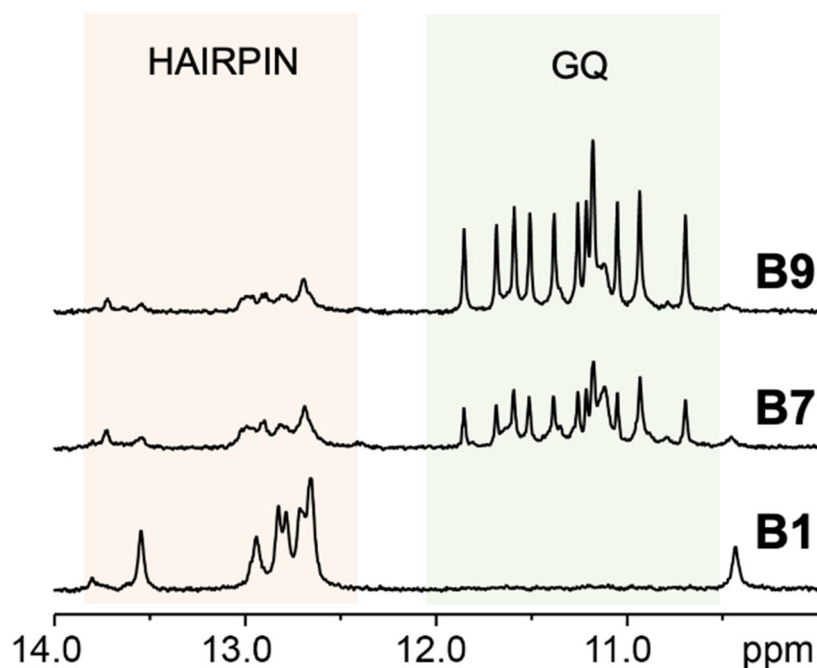


Figure 3. ^1H NMR spectra (imino region) of CEBm4 under different ionic conditions (B1: 140 mM Na^+ / 0 mM K^+ ; B7: 70 mM Na^+ / 70 mM K^+ ; B9: 0 mM Na^+ / 140 mM K^+) in 10 mM LiCacO buffer, pH = 7.2.

B1 and B9 buffers are close to extra- and intra-cellular ionic conditions, respectively. We used the peroxidase-like G4-DNAzyme activity assay to confirm the conformation transition of m4. Since the G4/hemin complex can catalyze the oxidation of ABTS^{2-} [37] in the presence of H_2O_2 , formation of a G4 structure could be evidenced by the accumulation of the chromogenic oxidation product $\text{ABTS}^{\bullet-}$, which has specific absorbance properties. As shown in **Figure 4**, CEBwt showed good catalytic property under nearly all buffer conditions. In contrast, CEBm7 was completely unable to catalyze the oxidation reaction of ABTS^{2-} even in B9 buffer, implying that m7 was fully folded into a duplex in all buffer conditions, in agreement with UV-melting results. CEBm4 is not able to catalyze the oxidation of ABTS^{2-} in the absence of potassium (B1), since $\text{ABTS}^{\bullet-}$ absorbance at 420 nm is close to the one of CEBm7. Differently, the catalytic efficiency of the DNAzyme increased in B2-B4 buffers to reach levels comparable to CEBwt in B7-B9 buffers containing high $[\text{K}^+]$, implying that quadruplexes may be partially or predominantly formed

under these conditions. Hence, the DNAzyme activity confirms that the structure of CEBm4 can be converted from hairpin to quadruplex by changing the potassium / sodium ratio.

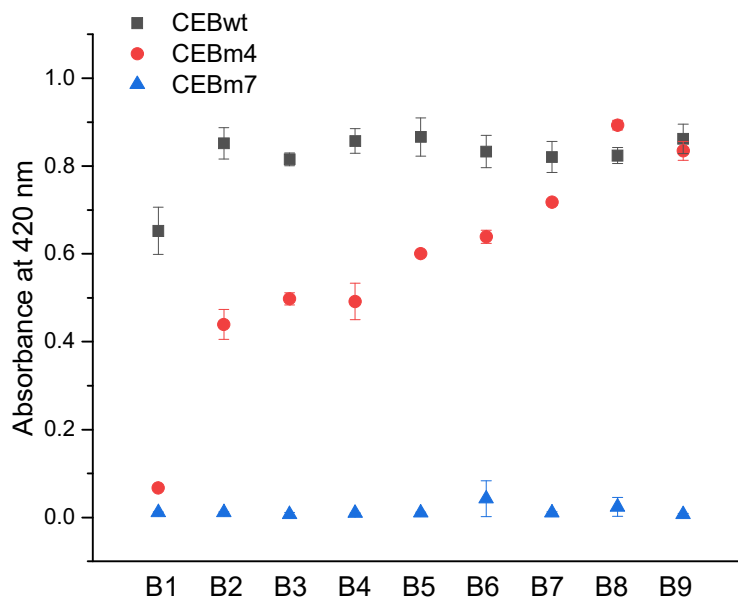


Figure 4. Absorbance at 420 nm resulting from H_2O_2 -dependent oxidation of $ABTS^{2-}$ to $ABTS^{\bullet -}$ by three different sequences (CEBwt, CEBm4 and CEBm7) in B1-B9 buffers.

N-methyl mesoporphyrin IX (NMM) is a parallel G4-specific fluorescence dye. In the presence of CEBm4, the fluorescence of NMM does not change significantly in B1 as compared to the negative control (CEBm7), suggesting that CEBm4 is barely forming a G4 in these conditions. Conversely, the fluorescence signal raised in B2 buffer and this increase is even more pronounced in B3-B9 buffers containing higher $[K^+]$, implying that quadruplexes exist under these conditions (**Figure S6**).

The evidence shown in § 3.3 demonstrated that the structures of CEBm4 in B1 and B9 are completely different: CEBm4 mainly folded into a hairpin in B1 and in a quadruplex in B9. Additionally, DNAzyme activity and NMM staining results suggest that quadruplexes may be partially folded in low potassium concentration buffer (B2-B4), while G4 structure traces in these potassium-deficient environments could not be followed by UV-melting (**Figure S4e**). There are two possible reasons explaining these discrepancies: *i*) UV absorbance at 295 nm may be too weak to be followed, as the preferred conformation of CEBm4 in potassium-deficient buffers is a hairpin, rather than a quadruplex; *ii*) Both hemin and NMM are G4 ligands,

they may induce to forming G4 structure and ‘rise up’ G4 proportion. In another word, G4 ligands may shift the hairpin/G4 equilibrium toward quadruplex.

4. Discussion and conclusion

The T_m of the CEB sequences measured in different buffers revealed that the presence of cytosines in G4 loops does not decrease G4s stability, but rather increases the probability of forming a competing structure, either an intramolecular hairpin or a duplex. In detail, the presence of three continuous cytosines in the central loop barely influences G4 formation under all potassium / sodium conditions tested. With an increasing number of cytosines, the duplex starts to be the predominant structure under low potassium concentrations (0 - 40 mM), while G4 folding still occurs in the presence of sufficient potassium ion concentration (40 - 140 mM). The T_m values of these G4 structures only depend on the specific potassium / sodium ratio, rather than on the specific cytosine-content of the mutated sequences. When the number of ‘CCC’ runs further increases to three, the CEB mutated sequence predominately folds into a duplex under all ionic conditions. Differently, T_m values of duplexes are barely affected by changes in $[K^+] / [Na^+]$ ratios and only depend on the primary sequence (GC content). These properties are summarized in the three schematic phase diagrams shown in **Figure 5**.

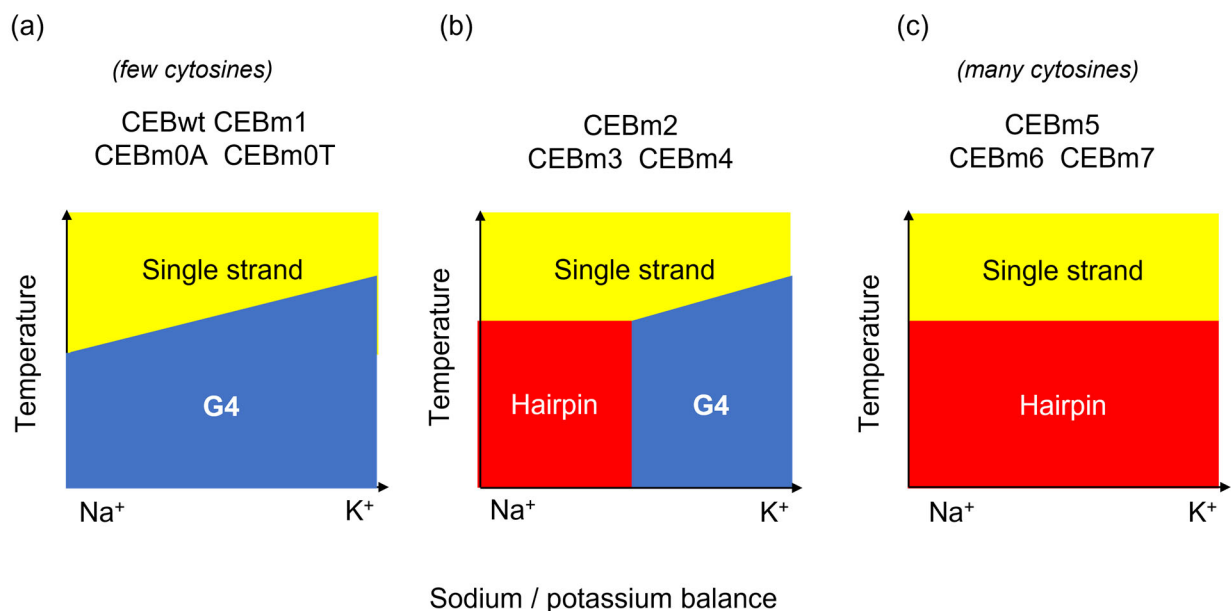


Figure 5. Phase diagrams showing the predominant species (single-strand, hairpin or G4) as a function of temperature and sodium / potassium ratio. 3 situations were encountered: (a) Very few cytosines are

present in the primary sequence, and the quadruplex always predominates at lower temperatures; T_m depends on K^+ concentration. (b) Intermediate case, in which the presence of a limited number of CCC runs allows the hairpin to predominate at low temperature and low potassium concentration, while a G4 is formed once a certain K^+ concentration threshold is reached (exact threshold being sequence-dependent). (c) C-rich sequences do not adopt a quadruplex fold, no matter the concentration of potassium: the hairpin predominates at low temperature; its T_m depends on exact sequence but is independent on K^+ concentration.

References

1. M. Moustakas, *The Role of Metal Ions in Biology, Biochemistry and Medicine*. Materials (Basel), 2021. **14**(3).
2. A.K. Rajasekaran and S.A. Rajasekaran, *Role of Na-K-ATPase in the assembly of tight junctions*. American Journal of Physiology-Renal Physiology, 2003. **285**(3): p. F388-F396.
3. V. Perez and E.T. Chang, *Sodium-to-potassium ratio and blood pressure, hypertension, and related factors*. Adv Nutr, 2014. **5**(6): p. 712-41.
4. V.M. Vitvitsky, S.K. Garg, R.F. Keep, R.L. Albin, and R. Banerjee, *Na⁺ and K⁺ ion imbalances in Alzheimer's disease*. Biochim Biophys Acta, 2012. **1822**(11): p. 1671-81.
5. A.L. Valton, V. Hassan-Zadeh, I. Lema, N. Boggetto, P. Alberti, C. Saintome, J.F. Riou, and M.N. Prioleau, *G4 motifs affect origin positioning and efficiency in two vertebrate replicators*. EMBO J, 2014. **33**(7): p. 732-46.
6. D. Sun and L.H. Hurley, *The Importance of Negative Superhelicity in Inducing the Formation of G-Quadruplex and i-Motif Structures in the c-Myc Promoter: Implications for Drug Targeting and Control of Gene Expression*. Journal of Medicinal Chemistry, 2009. **52**(9): p. 2863-2874.
7. C. Weldon, J.G. Dacanay, V. Gokhale, P.V.L. Boddupally, I. Behm-Ansmant, G.A. Burley, C. Branlant, L.H. Hurley, C. Dominguez, and I.C. Eperon, *Specific G-quadruplex ligands modulate the alternative splicing of Bcl-X*. Nucleic Acids Res, 2018. **46**(2): p. 886-896.
8. E. Largy, A. Marchand, S. Amrane, V. Gabelica, and J.L. Mergny, *Quadruplex Turncoats: Cation-Dependent Folding and Stability of Quadruplex-DNA Double Switches*. J Am Chem Soc, 2016. **138**(8): p. 2780-92.
9. H. Tateishi-Karimata, K. Kawauchi, and N. Sugimoto, *Destabilization of DNA G-Quadruplexes by Chemical Environment Changes during Tumor Progression Facilitates Transcription*. J Am Chem Soc, 2018. **140**(2): p. 642-651.
10. G. Biffi, D. Tannahill, J. Miller, W.J. Howat, and S. Balasubramanian, *Elevated levels of G-quadruplex formation in human stomach and liver cancer tissues*. PLoS One, 2014. **9**(7): p. e102711.
11. M. Chen, Q. Chen, Y. Li, Z. Yang, E.W. Taylor, and L. Zhao, *A G-quadruplex nanoswitch in the SGK1 promoter regulates isoform expression by K(+)/Na(+) balance and resveratrol binding*. Biochim Biophys Acta Gen Subj, 2021. **1865**(2): p. 129778.
12. S. Amrane, M. Adrian, B. Heddi, A. Serero, A. Nicolas, J.L. Mergny, and A.T. Phan, *Formation of pearl-necklace monomeric G-quadruplexes in the human CEB25 minisatellite*. J Am Chem Soc, 2012. **134**(13): p. 5807-16.
13. A.K. Todd, M. Johnston, and S. Neidle, *Highly prevalent putative quadruplex sequence motifs in human DNA*. Nucleic Acids Res, 2005. **33**(9): p. 2901-7.
14. J. Marquevielle, A. De Rache, B. Vialet, E. Morvan, J.L. Mergny, and S. Amrane, *G-quadruplex structure of the C. elegans telomeric repeat: a two tetrads basket type conformation stabilized by a non-canonical C-T base-pair*. Nucleic Acids Res, 2022.
15. V. Brazda, Y. Luo, M. Bartas, P. Kaura, O. Porubiakova, J. Stastny, P. Pecinka, D. Verga, V. Da Cunha, T.S. Takahashi, P. Forterre, H. Myllykallio, M. Fojta, and J.L. Mergny, *G-Quadruplexes in the Archaea Domain*. Biomolecules, 2020. **10**(9).

16. A. Cantara, Y. Luo, M. Dobrovolna, N. Bohalova, M. Fojta, D. Verga, L. Guittat, A. Cucchiarini, S. Savrimoutou, C. Haberli, J. Guillon, J. Keiser, V. Brazda, and J.L. Mergny, *G-quadruplexes in helminth parasites*. *Nucleic Acids Res*, 2022. **50**(5): p. 2719-2735.
17. M. Dobrovolná, N. Bohálová, V. Peška, J. Wang, Y. Luo, M. Bartas, A. Volná, J.-L. Mergny, and V. Brázda, *The Newly Sequenced Genome of Pisum sativum Is Replete with Potential G-Quadruplex-Forming Sequences—Implications for Evolution and Biological Regulation*. *International Journal of Molecular Sciences*, 2022. **23**(15).
18. A. Bedrat, L. Lacroix, and J.L. Mergny, *Re-evaluation of G-quadruplex propensity with G4Hunter*. *Nucleic Acids Res*, 2016. **44**(4): p. 1746-59.
19. K.W. Zheng, J.Y. Zhang, Y.D. He, J.Y. Gong, C.J. Wen, J.N. Chen, Y.H. Hao, Y. Zhao, and Z. Tan, *Detection of genomic G-quadruplexes in living cells using a small artificial protein*. *Nucleic Acids Res*, 2020. **48**(20): p. 11706-11720.
20. A. Guedin, J. Gros, P. Alberti, and J.L. Mergny, *How long is too long? Effects of loop size on G-quadruplex stability*. *Nucleic Acids Res*, 2010. **38**(21): p. 7858-68.
21. V.T. Mukundan and A.T. Phan, *Bulges in G-Quadruplexes: Broadening the Definition of G-Quadruplex-Forming Sequences*. *Journal of the American Chemical Society*, 2013. **135**(13): p. 5017-5028.
22. X.M. Li, K.W. Zheng, J.Y. Zhang, H.H. Liu, Y.D. He, B.F. Yuan, Y.H. Hao, and Z. Tan, *Guanine-vacancy-bearing G-quadruplexes responsive to guanine derivatives*. *Proc Natl Acad Sci U S A*, 2015. **112**(47): p. 14581-6.
23. J.L. Mergny, J. Li, L. Lacroix, S. Amrane, and J.B. Chaires, *Thermal difference spectra: a specific signature for nucleic acid structures*. *Nucleic Acids Res*, 2005. **33**(16): p. e138.
24. E. Puig Lombardi, A. Holmes, D. Verga, M.P. Teulade-Fichou, A. Nicolas, and A. Londono-Vallejo, *Thermodynamically stable and genetically unstable G-quadruplexes are depleted in genomes across species*. *Nucleic Acids Res*, 2019. **47**(12): p. 6098-6113.
25. A. Guédin, A. De Cian, J. Gros, L. Lacroix, and J.-L. Mergny, *Sequence effects in single-base loops for quadruplexes*. *Biochimie*, 2008. **90**(5): p. 686-696.
26. S. Amrane, M. Adrian, B. Heddi, A. Serero, A. Nicolas, J.-L. Mergny, and A.T. Phan, *Formation of Pearl-Necklace Monomorphic G-Quadruplexes in the Human CEB25 Minisatellite*. *Journal of the American Chemical Society*, 2012. **134**(13): p. 5807-5816.
27. L. Trantírek, R. Štefl, M. Vorlíčková, J. Koča, V.r. Sklenář, and J. Kypr, *An A-type double helix of DNA having B-type puckering of the deoxyribose rings* | Edited by I. Tinoco. *Journal of Molecular Biology*, 2000. **297**(4): p. 907-922.
28. J. Kypr, I. Kejnovska, D. Renciuik, and M. Vorlickova, *Circular dichroism and conformational polymorphism of DNA*. *Nucleic Acids Res*, 2009. **37**(6): p. 1713-25.
29. G. Ma, Z. Yu, W. Zhou, Y. Li, L. Fan, and X. Li, *Investigation of Na⁺ and K⁺ Competitively Binding with a G-Quadruplex and Discovery of a Stable K⁺-Na⁺-Quadruplex*. *The Journal of Physical Chemistry B*, 2019. **123**(26): p. 5405-5411.
30. J. Chen, M. Cheng, P. Stadlbauer, J. Šponer, J.-L. Mergny, H. Ju, and J. Zhou, *Exploring Sequence Space to Design Controllable G-Quadruplex Topology Switches*. *CCS Chemistry*, 2021: p. 3232-3246.

31. P.V. Riccelli, K.E. Mandell, and A.S. Benight, *Melting studies of dangling-ended DNA hairpins: effects of end length, loop sequence and biotinylation of loop bases*. *Nucleic Acids Research*, 2002. **30**(18): p. 4088-4093.
32. S.V. Kuznetsov, Y. Shen, A.S. Benight, and A. Ansari, *A Semiflexible Polymer Model Applied to Loop Formation in DNA Hairpins*. *Biophysical Journal*, 2001. **81**(5): p. 2864-2875.
33. S.V. Kuznetsov, C.C. Ren, S.A. Woodson, and A. Ansari, *Loop dependence of the stability and dynamics of nucleic acid hairpins*. *Nucleic Acids Res*, 2008. **36**(4): p. 1098-112.
34. M.J.J. Blommers, J.A.L.I. Walters, C.A.G. Haasnoot, J.M.A. Aelen, G.A. Van der Marel, J.H. Van Boom, and C.W. Hilbers, *Effects of base sequence on the loop folding in DNA hairpins*. *Biochemistry*, 1989. **28**(18): p. 7491-7498.
35. M. Adrian, B. Heddi, and A.T. Phan, *NMR spectroscopy of G-quadruplexes*. *Methods*, 2012. **57**(1): p. 11-24.
36. S. Dzatko, M. Krafcikova, R. Hansel-Hertsch, T. Fessl, R. Fiala, T. Loja, D. Krafcik, J.L. Mergny, S. Foldynova-Trantirkova, and L. Trantirek, *Evaluation of the Stability of DNA i-Motifs in the Nuclei of Living Mammalian Cells*. *Angew Chem Int Ed Engl*, 2018. **57**(8): p. 2165-2169.
37. P. Travascio, Y. Li, and D. Sen, *DNA-enhanced peroxidase activity of a DNA aptamer-hemin complex*. *Chemistry & Biology*, 1998. **5**(9): p. 505-517.

A sodium / potassium switch for G4-prone GC-rich sequences

Supplementary Information
in preparation, December 18 version.

Yu Luo^{1,4}, Martina Lenarčič Živkovič^{2,3} (...)
Daniela Verga^{4,5*}, Lukáš Trantírek^{2*} & Jean-Louis Mergny^{1*}

1. Laboratoire d'Optique et Biosciences, Ecole Polytechnique, CNRS, Inserm, Institut Polytechnique de Paris, 91128 Palaiseau, France;
2. Central European Institute of Technology, Masaryk University, 625 00 Brno, Czech Republic;
3. Slovenian NMR Centre, National Institute of SI-1000 Chemistry Ljubljana, Slovenia;
4. CNRS UMR9187, INSERM U1196, Université Paris-Saclay, F-91405 Orsay, France;
5. CNRS UMR9187, INSERM U1196, Institut Curie, PSL Research University, F-91405 Orsay, France;

Contents :

- *Table S1*
- *Figures S1-S6*
- *Additional references*

Table S1. Numbers of PQS found in human genome (hg19) (Database: GSE133379 [1])

Chromosome	Nos of PQS	PQS per kbp	Nos of 'CCC' contained PQS	Frequency of 'CCC' in PQS	Nos of two 'CCC's contained PQS
Chr 1	134804	0.54	13779	0.10	1068
Chr 2	110685	0.46	10124	0.09	671
Chr 3	79202	0.40	6412	0.08	424
Chr 4	60491	0.32	5138	0.08	357
Chr 5	70948	0.39	5924	0.08	391
Chr 6	68177	0.40	5358	0.08	369
Chr 7	78345	0.49	7823	0.10	597
Chr 8	64136	0.44	5977	0.09	447
Chr 9	69477	0.49	7659	0.11	627
Chr 10	71686	0.53	7336	0.10	561
Chr 11	81053	0.60	8566	0.11	675
Chr 12	64896	0.48	6131	0.09	565
Chr 13	32452	0.28	2737	0.08	244
Chr 14	46353	0.43	4687	0.10	346
Chr 15	49090	0.48	4953	0.10	309
Chr 16	65359	0.72	7996	0.12	708
Chr 17	74609	0.92	8919	0.12	682
Chr 18	30684	0.39	2669	0.09	199
Chr 19	75540	1.28	10294	0.14	902
Chr 20	46930	0.74	5162	0.11	440
Chr 21	19109	0.40	2381	0.12	226
Chr 22	43195	0.84	5930	0.14	461
Chr X	59843	0.39	4384	0.07	289
Chr Y	9289	0.16	781	0.08	48
Sum	1506353	-	151120	-	11606

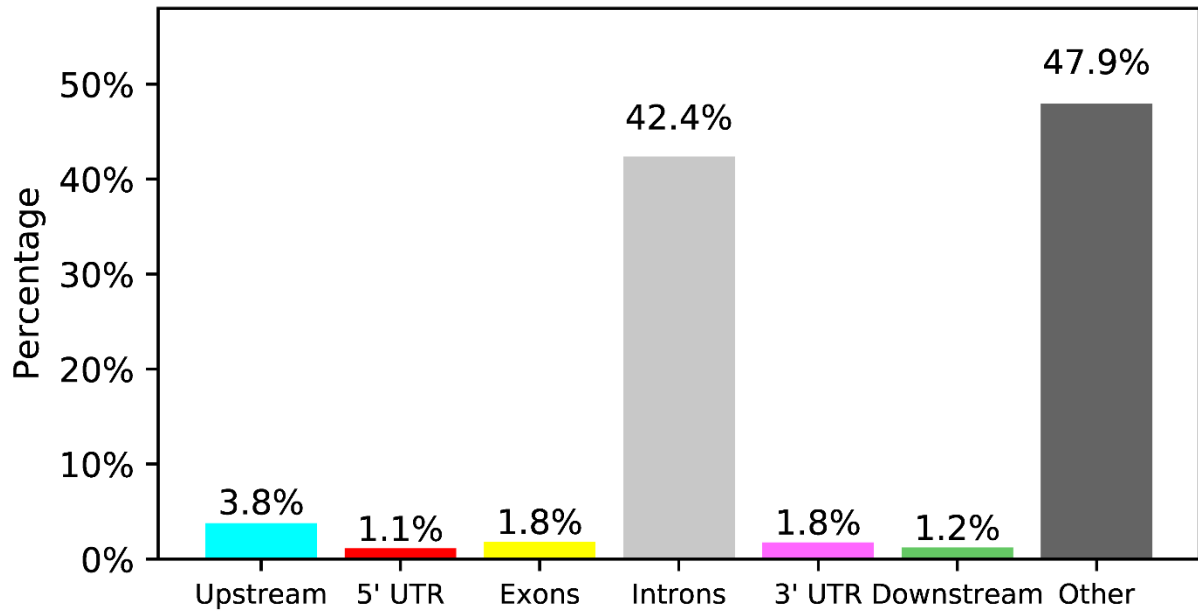


Figure S1. Genomic location and distribution of PQS containing a single 'CCC'-motif.

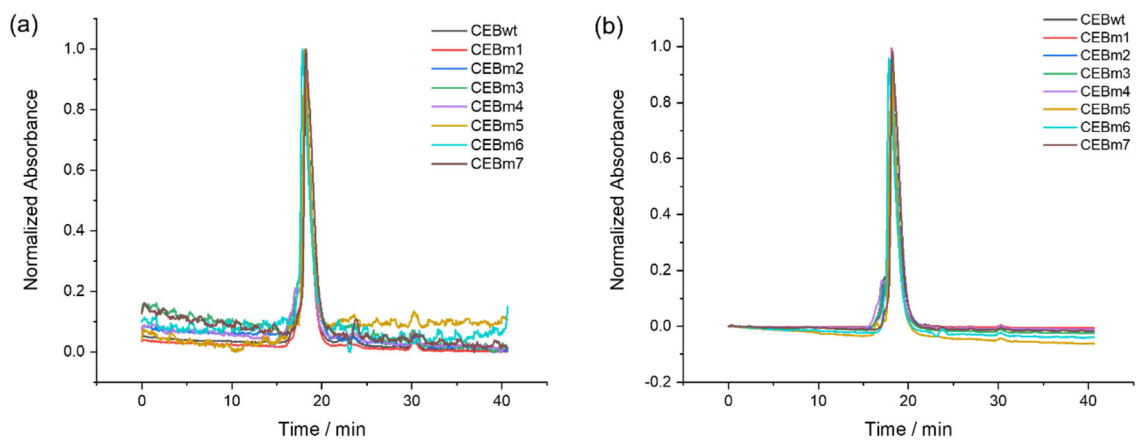


Figure S2. Analysis of the folded species according to SE-HPLC analysis. Normalized absorbance of mutations at (a) 10 mM LiCaco (pH = 7.2) buffer with 140 mM NaCl and (b) 10 mM LiCaco (pH = 7.2) buffer with 140 mM KCl. A single predominant species is found for all sequences, matching the peak obtained for CEB25wt, previously determined to be a monomolecular species [2].

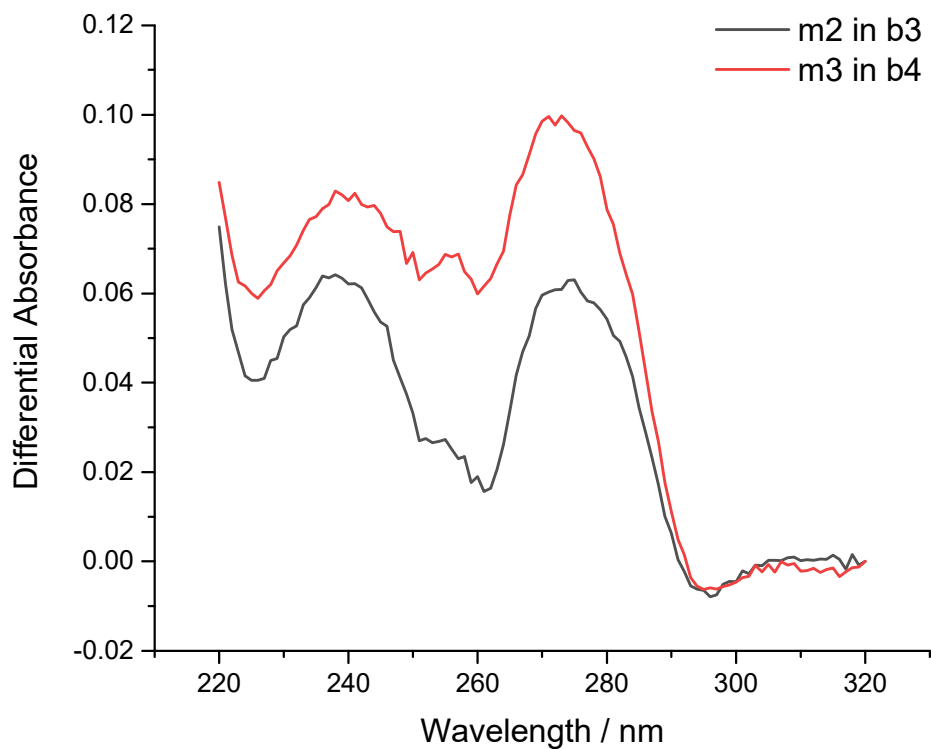


Figure S3. Thermal differential spectra (TDS) of 5 μM CEBm2 in B3 buffer (125 mM NaCl, 15 mM KCl in 10 mM LICaco, pH = 7.2) and CEBm3 in B4 buffer (115 mM NaCl, 25 mM KCl in 10 mM LICaco, pH = 7.2). TDS corresponds to the arithmetic difference between the initial (25 $^{\circ}\text{C}$) and second (95 $^{\circ}\text{C}$) spectra.

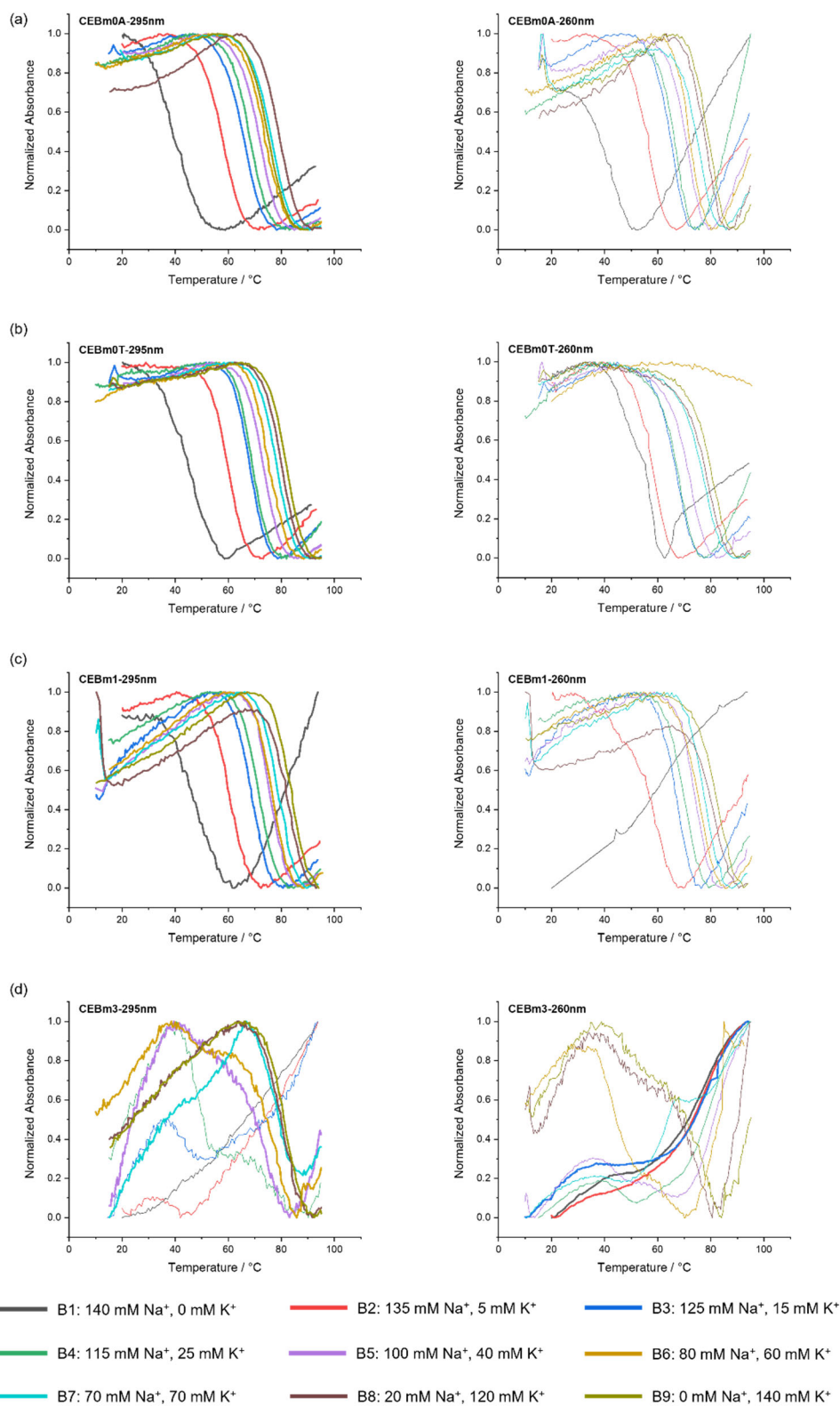


Figure S4. To be continued next page.

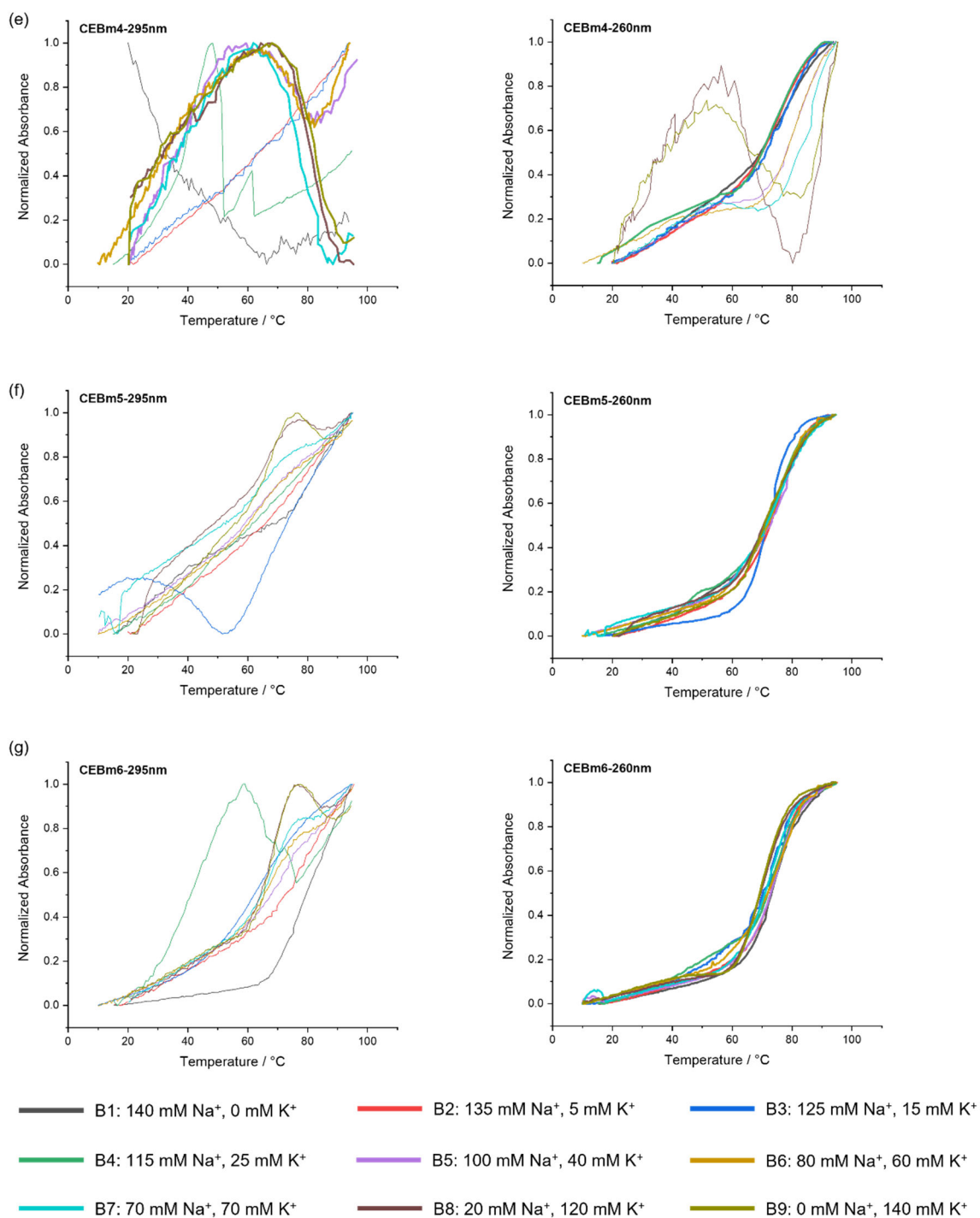


Figure S4. Normalized UV-melting curves of 5 μ M CEB mutations in different buffers. Heating runs were performed between 10 $^{\circ}$ C and 95 $^{\circ}$ C, and the temperature was increased by 0.2 $^{\circ}$ C per minute, and the absorbance was recorded at 260 and 295 nm. Curves shown in bold were used to calculate the T_m .

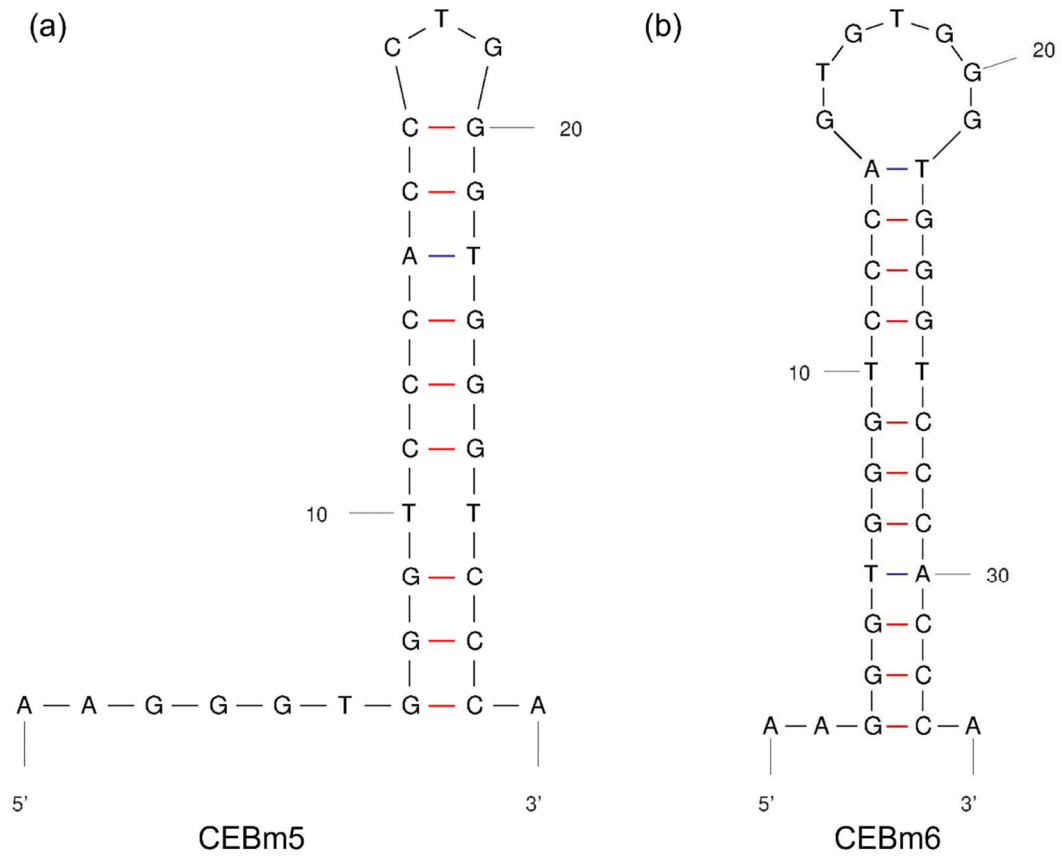


Figure S5. The hairpin models of CEBm5 and m6 according to UNAFold [3].

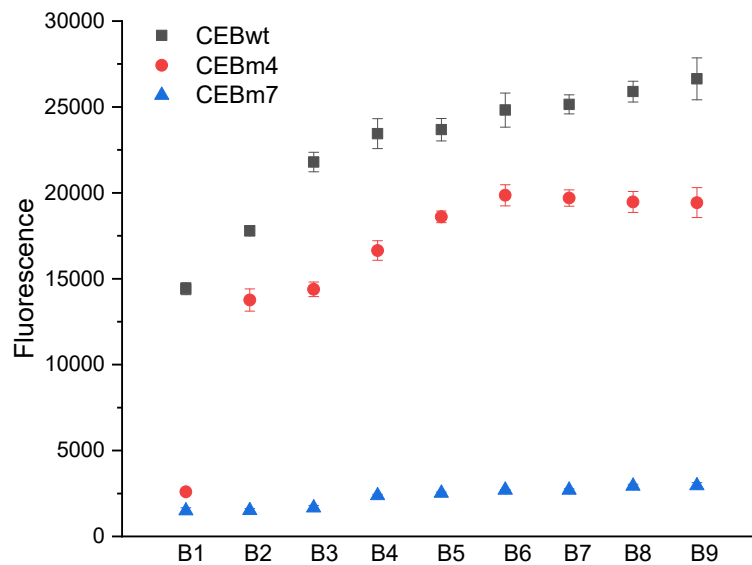


Figure S6 2 μM NMM staining of structures generated by 3 μM oligonucleotides in different buffers. Fluorescence was immediately measured after NMM addition.

Additional References:

1. K.W. Zheng, J.Y. Zhang, Y.D. He, J.Y. Gong, C.J. Wen, J.N. Chen, Y.H. Hao, Y. Zhao, and Z. Tan, *Detection of genomic G-quadruplexes in living cells using a small artificial protein*. *Nucleic Acids Res*, 2020. **48**(20): p. 11706-11720.
2. S. Amrane, M. Adrian, B. Heddi, A. Serero, A. Nicolas, J.-L. Mergny, and A.T. Phan, *Formation of Pearl-Necklace Monomorphic G-Quadruplexes in the Human CEB25 Minisatellite*. *Journal of the American Chemical Society*, 2012. **134**(13): p. 5807-5816.
3. N.R. Markham and M. Zuker, *UNAFold*, in *Bioinformatics: Structure, Function and Applications*, J.M. Keith, Editor. 2008, Humana Press: Totowa, NJ. p. 3-31.

Chapter VI. General conclusion and perspectives

A major purpose of this thesis was to devise high-throughput methods for G4 characterization *in vitro*. The T_m of an intramolecular dual-labeled G4 structure (fluorophore - G4 sequence - quencher) can be followed by FRET-melting. At low temperature, the fluorophore and the quencher located on two ends of the G4 sequence are in close proximity, and the fluorescence is quenched. During melting, G4 unfolds and the fluorophore and the quencher are sent far apart, and fluorescence is restored. FRET-melting can also be used to study the interaction between G4s and ligands: if a ligand is able to stabilize a G4 higher T_m values are measured. Based on this assay, we first developed the FRET-MC assay: in this experiment we challenged the binding of PhenDC3 to F21T by the presence of an unlabeled competitor. As the probe G4 strand we used the dual-labeled (by a donor chromophore (FAM) and an acceptor (TAMRA)) well-established F21T sequence. The interaction between F21T and PhenDC3 is challenged only when the competitor forms a G4 structure, resulting in a drop in F21T T_m value, from F21T-PhenDC3 to F21T-alone. Compared to traditional spectrometric methods, FRET-MC takes the advantage of the classical FRET-melting: it can be performed in 96 microwell plates, which means it allows to analyze 48 competitors within 2 hours (one melting process). From a cost-analysis perspective, FRET-MC only needs one labeled sequence (F21T), and the amount of F21T used for the assay is extremely low ($0.2 \mu\text{M} \times 25 \mu\text{L}$ per well). We validated the FRET-MC assay by using a training set of sequences, which includes more than sixty known sequences. We measured the ΔT_m of F21T in the presence of PhenDC3 and in the absence or the presence of a competitor belonging to the training set. ΔT_m were then normalized into *S Factor* parameter to quantitate the competition effect, the empirical boundaries of G4 and non G4 competitors [247] were identified: *i) S < 0.3: G4 competitor; ii) 0.3 ≤ S < 0.6: unknown; iii) S ≥ 0.6: non-G4 competitor.* FRET-MC has been used to validate several PQS, but it does not work well to characterize G4s with weak thermal stability.

For this reason, we developed an alternative isothermal FRET assay compatible with weakly stable G4s. Iso-FRET exploits two labeled RNA probe strands: one of them 37Q (37-quencher) can form a G4 thanks to the binding of PhenDC3, and the second one F22 (FAM-22) is partially complementary to 37Q. To this system an unlabeled competitor sequence is added. When the competitor is not a G4, PhenDC3 remains bound to 37Q and stabilizes its G4 structure, preventing the complex between 37Q and F22 to form, allowing a high fluorescence signal of F22. On the contrary, an excess of a G4 competitor traps PhenDC3, which is no longer available to bind to 37Q, which in turns allows this oligonucleotide to hybridize to F22 resulting in fluorescence quenching.

RNA sequences were used in this assay. Labeled RNA sequences are expensive but, as mentioned for FRET-MC, reaction volume and concentrations of the two probe strands are low (20 nM F22, 200 nM 37Q $\times 25 \mu\text{L}$ per well) meaning that only a tiny fraction of the synthesized nucleotides is used for each

point. We tried to design an alternative DNA-based assay, but adjusting the relative stabilities of the duplex and quadruplex turned out to be tricky, and results were not as accurate. For this reason, we kept the original labeled RNA strands. Thousands of assays may be performed with a single RNA chemically synthesized sequence. In addition to addressing the false negative problem associated to FRET-MC of G4 competitors with poor thermal stability, iso-FRET can be performed at room temperature (25 °C) and at physiological temperature (37 °C). Iso-FRET also simplifies data handling, and increases the throughput from 96 to 384 samples per plate. Similar to FRET-MC, the accuracy of iso-FRET assay was evidenced by the training set sequences; F value was defined to normalized the fluorescence intensity at 25 °C and it was used to calculate the boundaries in a 95% prediction interval: *i*) $F < 0.33$: G4 competitor; *ii*) $0.33 \leq F < 0.54$: unknown; *iii*) $F \geq 0.54$: non-G4 competitor.

Iso-FRET is still not a perfect assay for G4 characterization, as a potential pitfall was identified. A G-rich competitor showing high complementarity to F22 may form a competitor-F22 duplex, preventing fluorescence quenching and resulting in false negative results, even if this competitor is prone to G4 formation. To solve the problem, we quantitated the complementarity between the competitor and F22 into the CF factor, which was calculated by a simple alignment. The CF factor is between 0 to 1, where 0 means no complementarity between the competitor and F22, and 1 means the competitor is fully complementary to F22. Our results show that the iso-FRET assay is able to characterize DNA competitors with a CF factor lower than 0.86.

FRET-MC and iso-FRET are fast and reliable methods to characterize G4s *in vitro*. Both assays offer G4 characterization templates: PhenDC3 can be replaced by any other G4 topology-specific ligands, to get topological information in batch. The FRET-MC buffer can be changed depending on specific application requirements: for example, potassium in buffer could be replaced by sodium or other ions, and the concentration is also allowed to change. It should be noted that T_m of F21T at high potassium concentration may be too high to be determined and it may influence the accuracy of the assay. Cation strength in iso-FRET affects the competitive binding of the G4 ligands and the F22-37Q hybridization, and it is not easy to adjust ion concentrations in the buffer. Thanks to the simplified process of the iso-FRET, it is possible to combine the iso-FRET with programmed manipulator/robot, to increase the throughput to ten thousand sequences per day. Although not tested, we believe that other elements, such as the presence of crowding agents in the buffer, are compatible both with FRET-MC and iso-FRET.

G4 characterization is an interesting and broad topic and how to increase the throughput to validate predicted PQS is still a question. G4 scoring algorithms can be used to exclude manifestly unlikely PQS before wet experiments and reduce the workload in laboratory. G4Hunter is a great G4-predicting algorithm which has been used to map and score PQS in several genomes. G4Hunter tends to exclude sequences containing alternative runs of G and C (for example, $(GGGCCC)_n$ gets a G4H score of 0).

We wanted to investigate whether the penalty imposed by G4Hunter to the presence of C in a given G-

rich sequence was experimentally justified. Our bioinformatic results showed that about 10.0% of PQS in the human genome (hg19) comprised at least three continuous cytosines, which triggered our interest to study the effect produced by the presence of C-tracks in G-rich sequences. We determined that the three-quartet G4 structure (CEB25) tolerates two "CCC" tracks in potassium / sodium mixed buffers containing 40 mM or higher KCl concentration, implying that cytosine contained G4-forming sequences can still adopt a G4 structure in the intracellular environment, and for potassium concentrations down to 40 mM, although the G4Hunter score of CEBm4 is only 0.69. It therefore appears that the penalty imposed by the central CCCTCCC run (-9 for each of the CCC motifs) is a bit excessive; later versions of G4Hunter may be more accurate if we decrease the negative weight of cytosines. This will, however, come with a cost: the G4Hunter current version is perfectly "symmetrical": if a strand has a score of +1, its complementary strand gets -1. Any deviation from this will mean that two independent scores need to be calculated for the two strands of a double-helix.

In any case, and whatever the quantitative extent of the penalty calculated for the presence of cytosines, one should understand that this penalty is not introduced because of a detrimental effect of C on G4 stability: melting experiments of G4-forming sequences in different buffers revealed that cytosines in G4 loops do not prevent decrease G4s thermal stability [253, 254]. This penalty rather reflects an increased tendency of these sequences to adopt a competing hairpin formation. As bioinformatics analysis indicates that the majority of cytosine-contained PQS located at pre-mRNAs, it will be interesting to transpose our experiments on DNA sequences to RNA motifs, in order to determine if the latter is more or less prone to hairpin / G4 competition.

Appendix (articles not included in the main thesis manuscript)

G4 in Archaea. *G-Quadruplexes in the Archaea Domain*. Brázda V, **Luo Y**, Bartas M, Kaura P, Porubiaková O, Šťastný J, Pečinka P, Verga D, Da Cunha V, Takahashi TS, Forterre P, Myllykallio H, Fojta M, Mergny JL. *Biomolecules*. 2020 Sep 21;10(9):1349. doi: 10.3390/biom10091349

Archaea constitute one of the three great domains of the tree of life. As G4 DNA has been found in Bacteria and Eukarya, we assumed that G4 may also exist in Archaea. We first mapped PQS and studied their distributions in several typical archaea [92]. PQS richness in Archaea groups was found to be independent of evolutionary proximity. The location of PQS in archaeal genomes is not random: generally, non-protein-coding RNAs (tRNA, rRNA, and other ncRNA) have a greater density of G4-prone motifs than protein-coding genes. The PQS density in ncRNA is substantially above the average G4 density of the genome, while the mRNA PQS density is near to the average. Our work also illustrated the difference in PQS distributions among archaea species, which triggered research interests on specific archaeal PQS loci.

G4 in Helminths. *G-quadruplexes in helminth parasites*. Cantara A, **Luo Y**, Dobrovolná M, Bohalova N, Fojta M, Verga D, Guittat L, Cucchiari A, Savrimoutou S, Häberli C, Guillon J, Keiser J, Brázda V, Mergny JL. *Nucleic Acids Res*. 2022 Mar 21;50(5):2719-2735. doi: 10.1093/nar/gkac129.

We also studied PQS in helminth parasites and experimentally confirmed quadruplex formation for chosen motifs. We revealed that two G4 ligands (JG1057 and JG1352) had potent activity both against larval and adult stages of *Schistosoma mansoni* [93]. The mechanisms of G4 ligands work as antiparasitic drugs need to be further studied. These results therefore open new perspectives to develop therapeutic strategies against helminth parasites based on G4-forming motifs in their genes.









G4 in Plants. *The Newly Sequenced Genome of Pisum sativum Is Replete with Potential G-Quadruplex-Forming Sequences-Implications for Evolution and Biological Regulation*. Dobrovolná M, Bohálová N, Peška V, Wang J, **Luo Y**, Bartas M, Volná A, Mergny JL, Brázda V. *Int. J. Mol. Sci*. 2022 Jul 30;23(15):8482. doi: 10.3390/ijms23158482.

We analyzed quadruplex formation in candidate sequences from the green pea genome.

I mainly contributed in the G4 structure validation experiments in these works. Some traditional spectroscopy-based characteristic techniques (differential UV spectra and CD spectra) and FRET-MC were used to validate the G4 formation of representative PQS, identified by bioinformatic results.

Article

G-Quadruplexes in the Archaea Domain

Václav Brázda ^{1,*}, Yu Luo ², Martin Bartas ³, Patrik Kaura ⁴, Otilia Porubiaková ^{1,5}, Jiří Štastný ^{4,6}, Petr Pečinka ³, Daniela Verga ², Violette Da Cunha ⁷, Tomio S. Takahashi ⁷, Patrick Forterre ⁷, Hannu Myllykallio ⁸, Miroslav Fojta ¹ and Jean-Louis Mergny ^{1,8,*}

¹ Institute of Biophysics of the Czech Academy of Sciences, Královopolská 135, 612 65 Brno, Czech Republic; o.porubiakova@gmail.com (O.P.); fojta@ibp.cz (M.F.)

² Institut Curie, CNRS UMR9187, INSERM U1196, Université Paris Saclay, 91400 Orsay, France; yu.luo@curie.fr (Y.L.); Daniela.Verga@curie.fr (D.V.)

³ Department of Biology and Ecology/Institute of Environmental Technologies, Faculty of Science, University of Ostrava, 710 00 Ostrava, Czech Republic; dutartas@gmail.com (M.B.); petr.pecinka@osu.cz (P.P.)

⁴ Faculty of Mechanical Engineering, Brno University of Technology, Technická 2896/2, 616 69 Brno, Czech Republic; 160702@vutbr.cz (P.K.); stastny@fme.vutbr.cz (J.Š.)

⁵ Faculty of Chemistry, Brno University of Technology, Purkyňova 464/118, 612 00 Brno, Czech Republic

⁶ Mendel University in Brno, Zemědělská 1, 613 00 Brno, Czech Republic

⁷ Institut de Biologie Intégrative de la Cellule (I2BC), CNRS, Université Paris-Saclay, CEDEX, 91198 Gif-sur-Yvette, France; violette.da.cunha.vdc@gmail.com (V.D.C.); tomio.takahashi@i2bc.paris-saclay.fr (T.S.T.); patrick.forterre@pasteur.fr (P.F.)

⁸ Laboratoire d'Optique et Biosciences, Ecole Polytechnique, CNRS, INSERM, Institut Polytechnique de Paris, 91128 Palaiseau, France; hannu.myllykallio@polytechnique.edu

* Correspondence: vaclav@ibp.cz (V.B.); jean-louis.mergny@inserm.fr (J.-L.M.); Tel.: +42-05-4151-7231 (V.B.); Fax: +42-05-4121-1293 (V.B.)

Received: 11 August 2020; Accepted: 18 September 2020; Published: 21 September 2020



Abstract: The importance of unusual DNA structures in the regulation of basic cellular processes is an emerging field of research. Amongst local non-B DNA structures, G-quadruplexes (G4s) have gained in popularity during the last decade, and their presence and functional relevance at the DNA and RNA level has been demonstrated in a number of viral, bacterial, and eukaryotic genomes, including humans. Here, we performed the first systematic search of G4-forming sequences in all archaeal genomes available in the NCBI database. In this article, we investigate the presence and locations of G-quadruplex forming sequences using the G4Hunter algorithm. G-quadruplex-prone sequences were identified in all archaeal species, with highly significant differences in frequency, from 0.037 to 15.31 potential quadruplex sequences per kb. While G4 forming sequences were extremely abundant in *Hadesarchaea archeon* (strikingly, more than 50% of the *Hadesarchaea archeon* isolate WYZ-LMO6 genome is a potential part of a G4-motif), they were very rare in the *Parvarchaeota* phylum. The presence of G-quadruplex forming sequences does not follow a random distribution with an over-representation in non-coding RNA, suggesting possible roles for ncRNA regulation. These data illustrate the unique and non-random localization of G-quadruplexes in Archaea.

Keywords: G4-forming motif; genome analysis; Archaea; unusual nucleic acid structures; sequence prediction

1. Introduction

The Archaea domain was classified separately from Bacteria by Carl Woese and George Fox in 1977 [1]. Later on, it was found that all major molecular machinery, such as DNA replication, transcription, and translation, of archaea are much more similar to those of eukaryotes than to those of

bacteria [2,3]. This is also true for some important membrane proteins, such as ATP synthases and proteins of the Sec transport system [4,5], or for some proteins involved in cell division and vesicle trafficking [6]. Thus, the archaeal domain occupies a key position in the Tree of Life, and there is currently a hot debate about their exact relationships with eukaryotes [7,8]. A schematic phylogenetic tree for the Archaea domain is proposed in Figure 1; this phylogeny is rapidly evolving with many new phyla recently identified via the accumulation of metagenome associated genomes (MAGs) and various new proposals for phylum definition and nomenclature [9,10]. The first detected archaea were isolated in harsh environments but later found in almost every environment, including the human microbiota, where they play important roles in the gut, mouth, and on the skin [11,12]. It has been hypothesized that archaea found in oceans are one of the most abundant groups of organisms on the planet with important roles both in the carbon and the nitrogen cycle [13]. The Archaea domain has several unique features, such as *ether*-linked lipids, while eukaryotes and most of the bacteria have ester-linked lipids [14]. Moreover, the stereochemistry of archaeal lipids has the opposite configuration as compare to the ones of eukaryotic and bacterial origin. Interestingly, methanogenesis, the production of greenhouse methane gas as a metabolic by-product, occurs only in the archaeal domain [15,16].

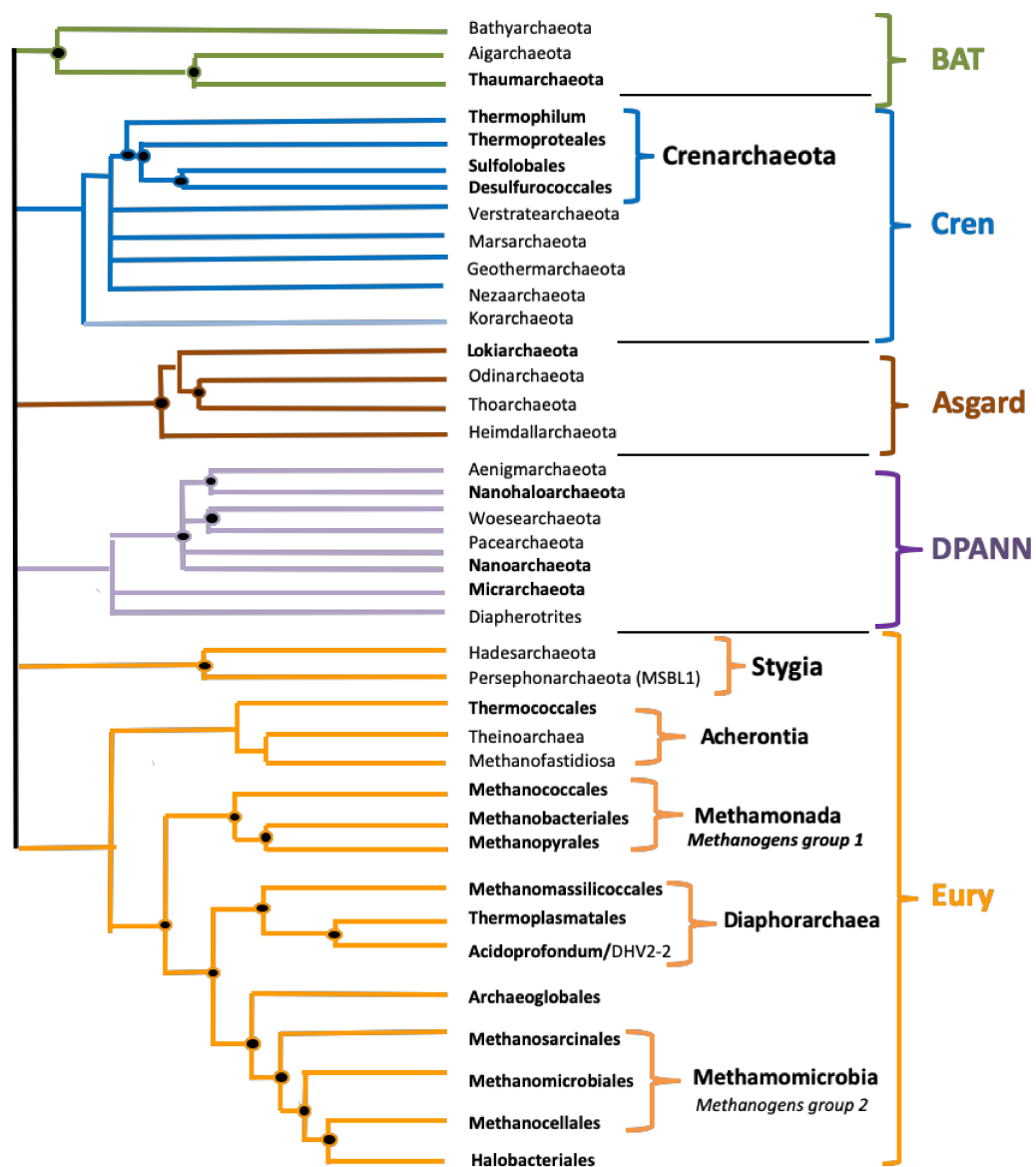


Figure 1. A schematic phylogenetic tree for Archaea. This unrooted evolutionary tree of Archaea is based

on the schematic tree of Forterre (2015) [17] updated according to recent phylogenetic analyses [9,18]. BAT stands for Bathyarchaeota, Aigarchaeota, and Thaumarchaeota. DPANN is an acronym based on the first five groups discovered: *Diapherotrites*, *Parvoarchaeota*, *Aenigmarchaeota*, *Nanoarchaeota*, and *Nanohaloarchaeota*. The term BAT superphylum has been proposed by Gaia et al. in 2018 [19], and the terms Eury and Cren superphyla are suggested here. The terms Cren superphylum is suggested here because the phyla *Crenarchaeota*, *Verstratearchaeota*, *Marsarchaeota*, *Nezaarchaeota*, and *Geothermarchaeota* form a consensus monophyletic clade in all archaeal phylogeny. We included *Korarchaeota* in this superphylum because they often branch as sister groups of the above phyla in archaeal phylogenies, although the fast evolutionary rate made their positioning sometimes difficult. We suggested in parallel the term Eury superphylum because Euryarchaeota includes very diverse groups of cultivated and uncultivated Archaea which are difficult to the group in a single phylum, especially considering that phyla, such as *Verstratearchaeota*, *Marsarchaeota*, or *Nezaarchaeota* only contain few uncultivated species only defined by a few metagenome associated genomes (MAGs). Names in bold letters correspond to subgroups that include cultivated species; names in thin letters correspond to subgroups that include only MAGs.

G-quadruplex structures (G4) formed by guanine rich sequences are among the most intensively studied local DNA/RNA structures [20]. G4s are formed by G:C Hoogsteen base pairing in a guanine quartet, and their formation requires the presence of stabilizing cations, such as potassium [21] (Figure 2). In both bacteria and eukaryotes, G4 formation regulates various processes, including gene expression [22], protein translation [23], and proteolysis [24]. G4 have been identified in a number of pathogens, including viruses, eukaryotes (e.g., *Plasmodium falciparum*) [25,26] or prokaryotes (e.g., *Neisseria gonorrhoeae* [27], and *Mycobacterium tuberculosis*) [28,29]. Moreover, many G4-binding proteins are conserved in all organisms highlighting the importance of the G4 structure regulations [30], and novel G4 binding proteins have been identified, sharing the NIQI amino acid motif (RGRGRRGGSGGSGGRGRG) [31]. Specific helicases have been identified both in eukaryotes and bacteria to unfold these structures, which can be extremely stable and would be problematic for the transcription or replication of G-rich motifs (e.g., the Pif1 or RecQ family helicases) [32]. Recently, G4Hunter was successfully used for the prediction of G-quadruplex-forming sequences in all complete bacterial genomes [33]. These results showed that G-quadruplex-forming sequences are present in all species with the highest frequencies in some extremophiles. In contrast to RNA, there is no correlation between genomic DNA GC% in Archaea (and in Bacteria) and the optimal growth temperature. This is likely because DNA in vivo is topologically closed, and topologically closed DNA is stable at least up to 107 °C [34]. We therefore cannot anticipate a higher density of G4-prone motifs in thermophiles, due to a GC-bias. A comparison with Extremophiles in bacteria is interesting [35]. Ding et al. hypothesized that stress-resistant bacteria found in the Deinococcales may utilize putative quadruplex sequences (PQS) for gene regulatory purposes. An enrichment in prokaryote PQS has been found in thermophilic organisms [33] but also in organisms with resistance to other stress factors, such as radiation [36,37]; thus, a direct correlation between temperature and G4 presence is not supported by these findings. In addition, while bacteria in the Deinococcus-Thermus group are the most abundant for PQS, it is striking that the mostly thermophilic and hyperthermophilic bacteria in the Thermotogae phylum have one of the lowest PQS frequencies. Correlation among thermophiles and G4s, therefore, depends on the phylum (Gram-negative vs. Gram-positive bacteria).

Due to the roles of G4s in the regulation of basic cellular processes, it is important to identify their location in genomes. Several algorithms are available to predict G-quadruplex-forming sequences [38–41]. Among them, the G4Hunter application was developed to provide quantitative analyses giving a propensity score as an output [41], and the G4Hunter web tool allows effective and fast analyses of PQS in large datasets [42].

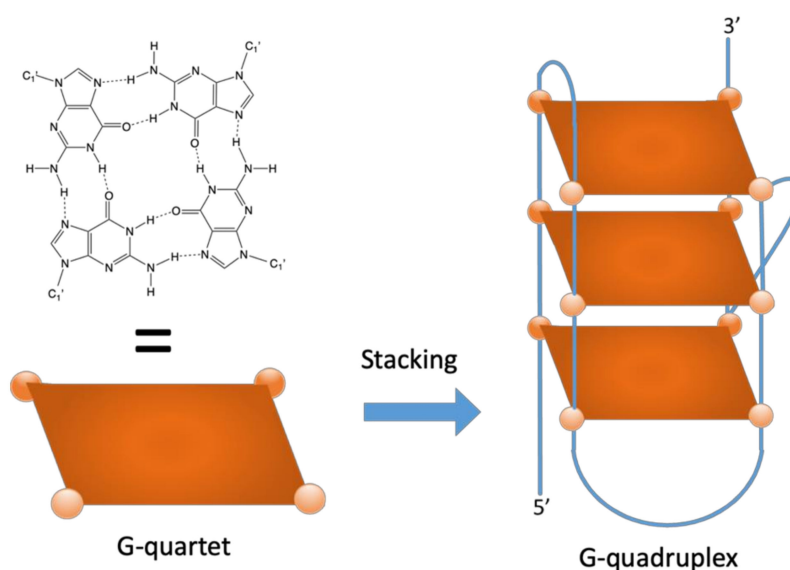


Figure 2. A G-quartet involves four coplanar guanines establishing a cyclic array of H-bonds (left). Stacking of two or more (three in this example) quartets leads to the formation of a G-quadruplex structure (right), stabilized by cations, such as potassium (not shown).

The prokaryotic genetic material is generally stored in circular chromosomes and plasmids [43]. The presence of quadruplex-prone motifs in over a hundred of bacterial genomes was determined over a decade ago [44]. In bacterial genomes, PQS are located non-randomly with a higher relative abundance in non-coding RNA (ncRNA), mRNA, and regions around tRNA and regulatory sequences. PQS also play roles in nitrate assimilation in *Paracoccus denitrificans* [45]. PQS in the *hsdS*, *recD*, and *pmrA* genes of *Streptococcus pneumoniae* contributes to host–pathogen interactions [46]. Such observations show the significant role of G4 in bacteria. The importance of another local DNA structure, the cruciform formed by inverted repeats, has been shown as an important regulatory feature of eukaryotic cell organelles, such as chloroplasts and mitochondria with circular DNA genomes [47,48]. Overall, the role of G4s in bacteria [27,49] and eukaryotes [50] is increasingly recognized.

In contrast, little is currently known regarding the abundance and location of PQS in the archaeal domain. Ding et al. performed an initial search on bacterial and archaeal genomes using a modified Quadparser algorithm with relaxed parameters allowing long loops (up to 12 nucleotides) [35]. They found that thermophilic microorganisms (both archaea and bacteria) appear to favor PQS in their genomes. Dhapola et al. created the Quadbase2 web server, in which G4 motifs found in a variety of organisms, including archaea, may be searched but did not analyze G4 propensity in archaea [51]. Because G4s play many important biological roles in bacterial and eukaryotic cells, we assume that G4s are also likely to have important functions in archaea. Therefore, we comprehensively analyzed the presence and locations of PQS in all sequenced archaeal genomes by G4Hunter [41,42]. These data provide the first study analyzing the presence of G4-prone sequences in this important domain of life.

2. Materials and Methods

2.1. Selection of the DNA Sequences

The set of all archaeal genomic DNA sequences was downloaded from the Genome database of the National Center for Biotechnology Information [52]. We have used for our analyses all accessible archaeal genomes, including contig and scaffold sequences (3387 genomes), and we have selected one representative genome for each species (Supplementary Table S1). For PQS analyses of features, we restricted our analysis to the subset of 140 completely assembled genomes. In total, we have

analyzed the presence of G4 forming sequences in 3387 genomes from the archaeal Domain representing a total of 6423 Mbps.

2.2. Process of Analysis

We used the computational core of our DNA analyzer software written in Java programming language [53]. For our analyses, we used a new G4Hunter algorithm implementation [42]. Default parameters for G4Hunter were set to “25” for window size and 1.2 or above for the G4H score (G4HS). PQS score was grouped to the five intervals: 1.2–1.4, 1.4–1.6, 1.6–1.8, 1.8–2.0, and 2.0 and more. Overall results for each species group contained a list of species with size of its genomic DNA sequence and number of putative G4 sequences found (Supplementary Table S2A); for clarity, the results for Groups and Subgroups are in separate files (Supplementary Table S2B,C). These data were processed by python jupyter using pandas with statistical tools [54]. Graphs were generated from the pandas tables using the “seaborn” graphical library. Note that the distinction between overlapping or discrete (non-overlapping) G4 motifs may create issues in the way potential motifs are counted. For this reason, we also provide a % PQS factor, which corresponds to the probability that any given nucleotide in the group or subgroup belongs to a G4-prone region ($G4H > 1.2$).

The default window value for G4Hunter has been discussed and tested in previous publications [41]. The value is chosen here (25 nt) corresponds more or less to the size of a typical intramolecular quadruplex. We considered shorter windows (20 nt) in previous studies. However, we noticed that for low thresholds (<1.2), a single GGGGGG run would give a hit; while intermolecular G4 formation is indeed possible with this motif, we hypothesized that intramolecular structures would be more relevant.

A slightly longer window (e.g., 30 nucleotides) further contributes to eliminating such motifs, but at the cost of significantly decreasing the number of hits (by a factor of 2 to 3; see Table 1): This larger window would, therefore, increase the number of false negatives, i.e., miss “real” intramolecular G4. On the other hand, a much larger window (50–100 nt) would be interesting to identify “G4 clusters” in which multiple tandem quadruplexes may be formed. We present the number of sequences found in three different complete archaeal genomes using four different window sizes and a threshold of 1.2:

Table 1. A number of putative quadruplex sequences (PQS) were found using four different window sizes in three complete archaeal genomes.

Archaea (GC %)	Number of G4 Sequences Found for a Window of:			
	25 nt	30 nt	50 nt	100 nt
<i>Methanococcus maripaludis</i> C7 (33.3%)	558	171	3	0
<i>Cenarchaeum symbiosum</i> A (57.3%)	6019	3197	324	5
<i>Halobacterium salinarum</i> NRC (65.9%)	4738	2313	262	4

As shown in Table 1, long G-rich prone regions, potentially supporting the formation of multiple quadruplexes, are present, but far less frequent (by a factor of 19 to 186 for a window of 50 vs. 25) than the classically defined G4Hunter motifs. In these three genomes, a large majority (95–99%) of the G4-prone regions would only support the formation of a single individual quadruplex.

2.3. Analysis of Putative G4 Sequences Around Annotated NCBI Features

We downloaded feature tables from the NCBI database along with genomic DNA sequences. Feature tables contain annotations of known features found in DNA sequences. We performed an analysis of G4-prone sequences occurrence inside recorded features. Features were grouped by their name stated in the feature table file (gene, rRNA, tRNA, ncRNA, and repeat region). From this analysis, we obtained a file with feature names and numbers of putative G4 forming sequences found inside and around features for each group of species analyzed. Search for putative G4 forming sequences

took place inside feature boundaries; note that frequencies of inverted repeats in mitochondrial DNA (mtDNA) [48], as well in the G4 prone sequences in bacteria [33], are distributed with different frequencies in close proximity to specific features. Further processing was performed in Microsoft Excel and the data are available as Supplementary Table S3.

2.4. Statistical Analysis

A cluster dendrogram of PQS characteristics was constructed in program R, version 3.6.3, library *pvclust* [55], to further reveal and graphically depict similarities between particular archaeal subgroups. Mean, Min, Max, and % PQS values were used as input data (Supplementary Table S4). The following parameters were used for analysis: Cluster method 'ward.D2', distance 'Euclidean', number of bootstrap resampling was set to 10,000. Statistically significant clusters (based on AU values (blue) above 95, equivalent to *p*-values less than 0.05) are highlighted by rectangles marked with broken red lines. R code is provided in Supplementary Table S4). Statistical evaluations of differences in G4 forming sequences presence in various phylogenetic groups were made by a Kruskal–Wallis test with a Bonferroni adjustment in STATISTICA, with *p*-value cut-off 0.05; data are available in Supplementary Table S5.

2.5. Quadruplex Formation In Vitro

Representative examples of the candidate sequences identified by G4Hunter were experimentally tested for G4 formation using different techniques: Isothermal difference spectra (IDS) and Circular dichroism (CD as described previously [41]).

2.5.1. Samples

Oligonucleotides were purchased from Eurogentec, Belgium, as dried samples purified by RP cartridge purification. Stock solutions were prepared at 250 μ M strand concentration in ddH₂O.

2.5.2. Experimental Conditions

Most experiments were performed in a 10 mM Lithium Cacodylate pH 7.1 buffer supplemented with 100 mM KCl (since *Hadesarchaea* has not been cultivated, it is impossible to know their intracellular potassium concentration. However, this is in the range of intracellular potassium concentration for other archaea, such as *Thermococcales*).

2.5.3. Isothermal Spectra

2.5 μ M oligonucleotide solutions were prepared in 10 mM Lithium Cacodylate buffer at pH 7.1. The solutions were kept at 95 °C for 5 min and slowly cooled to room temperature and kept at 4 °C overnight. Absorbance spectra were recorded on a Cary 300 (Agilent Technologies, France) spectrophotometer at 37 °C (scan range: 500–200 nm; scan rate: 600 nm/min; automatic baseline correction). After recording these first series of spectra (unfolded as no potassium was present) 1 M KCl (100 μ L) was added to the samples, and UV-absorbance spectra were recorded after 15 min equilibration, and corrected for dilution. Each IDS corresponds to the arithmetic difference between the initial (unfolded) and final (folded, corrected for dilution) spectra.

2.5.4. Circular Dichroism

2.5 μ M oligonucleotide solutions were prepared in 10 mM lithium cacodylate buffer at pH 7.1 supplemented with 100 mM KCl. The solutions were kept at 95 °C for 5 min and slowly cooled to room temperature and kept at 4 °C overnight. CD spectra were recorded on a JASCO J-1500 (France) spectropolarimeter at room temperature or at 80 °C, using a scan range of 400–210 nm, a scan rate of 200 nm/min, and averaging four accumulations (Supplementary Figure S1).

2.6. G-Quadruplex Binding Proteins Prediction

For G-quadruplex binding proteins prediction, based on previously published G-quadruplex binding motif (RGRGRGRGGGSGGSGGRGRG) [31], the BLASTp algorithm was used [56]. The target organisms were limited to the Archaea domain (NCBI taxid ID: 2157). E-value cut-off was set to 0.05. For similarity search of RecQ helicase from *Escherichia coli* (UNIPROT ID: P15043), BLASTp algorithm [56] was used with an E-value cut-off of 0.0001 and the same restriction to the Archaea domain, as above. BLASTp analyses are enclosed in Supplementary Table S6. FIMO search [57,58] for G-quadruplex binding motif (RGRGRGRGGGSGGSGGRGRG) [31] in *Methanosarcina mazei* complete proteome was carried out on a set of 15722 known protein sequences downloaded from NCBI, with q-value (*p*-value corrected for multiple testing by Benjamini and Hochberg method) cut-off of 0.05 (Supplementary Table S7). The most similar protein of RecQ helicase from *Escherichia coli* (UNIPROT ID: P15043) in *Hadesarchaea archaeon* isolate WYZ-LMO6 was searched using tBLASTn [59], and the resulting best hit was translated using ExPasy Translate Tool [60,61] and functional domain were visualized using NCBI CDD [62] (Supplementary Table S8).

3. Results

3.1. Prediction of G4 Forming Sequences in Archaea

We analyzed the occurrence of putative G4 sequences (PQS) with G4Hunter in 3387 archaeal genomes. The length of sequenced archaeal genomes in our dataset varied from 100 kbps to 13.4 Mbps (list provided in Supplementary Table S1). The average GC content was 46.51%, with a minimum of 24.30% for *Nanobsidianus stetteri* isolate SCGC AB-777 (*Nanoarchaeota*) and a maximum of 70.95% for *Halobacteriales archaeon* SW_7_71_33 (phylum *Euryarchaeota*). Using standard parameters for the G4Hunter search algorithm (window size of 25 and G4HS \geq 1.2) we found 4,470,813 PQS in these 3387 archaeal genomes using a default threshold of 1.2. The higher the G4HS score is, the higher the stability of the structure. Over 90% and 98% of sequences with a score above 1.2 or 1.5, respectively, were experimentally demonstrated to form a stable quadruplex in vitro [41]. Figure 3A provides an example of G-rich motifs found in archaea with G4HS between 1.32 and 3.0. As expected from previous analyses on eukaryotes and bacteria, most (97%) PQS have a relatively low (1.2 to 1.4) G4Hunter score. More stable motifs are rarer, with a sharp decrease in the number of retrieved sequences with scores above 1.4, as shown in Table 2. Only 132 PQS with a G4Hunter score of 2 or more were found. A summary of all PQS found in ranges of G4Hunter score intervals and precomputed PQS frequencies per 1000 bp is provided in Table 2.

Table 2. Number of PQS found and their frequencies per 1000 bp in all 3387 archaeal genomes, grouped by G4Hunter score (1.2-1.4 means any sequence with a score between 1.2 and 1.399; 1.4 between 1.4 and 1.599, etc.).

G4HS	Number of PQS in Dataset	Fraction of All PQS	PQS Frequency Per kbp
1.2–1.4	4,344,917	0.9718	1.19
1.4–1.6	119,233	0.0267	1.8×10^{-2}
1.6–1.8	6357	0.00142	9.9×10^{-4}
1.8–2.0	174	0.0000389	2.5×10^{-5}
>2.0	132	0.0000295	2.2×10^{-5}
Total	4,470,813	1	

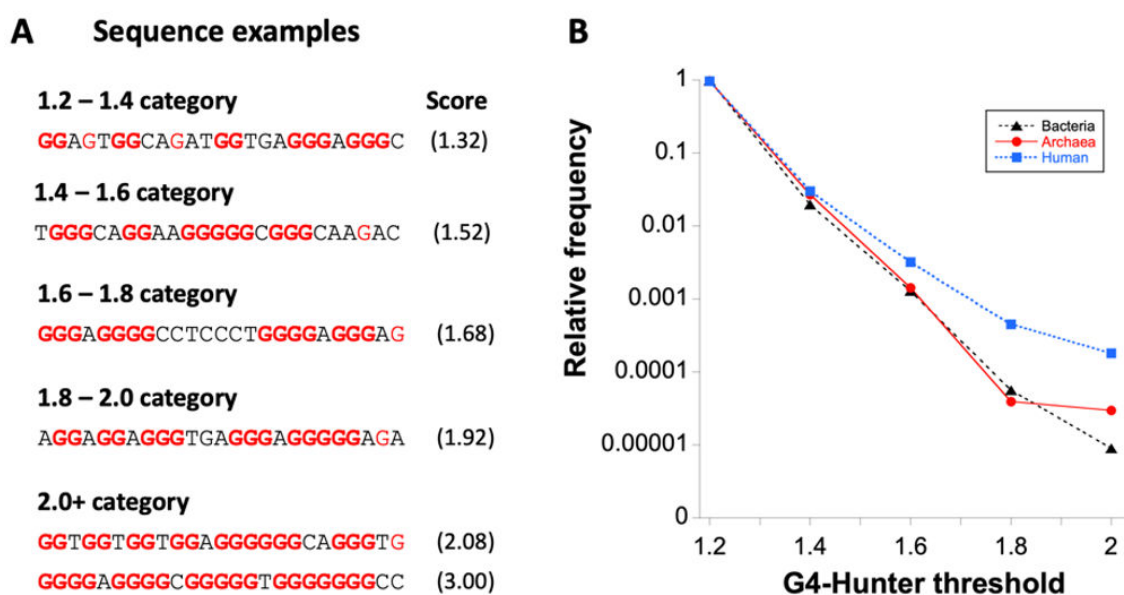


Figure 3. Examples of sequences with different G-quadruplexes (G4) Hunter scores (G4HS) and distribution of PQS according to threshold category. (A) Examples of archaea 25-nt long sequences (corresponding to the window size chosen for the analysis) for which G4Hunter scores are provided within parentheses. Isolated guanines are shown in red, all other guanines in bold red characters. Longer archaea motifs with high G4H scores are provided in Table 3. (B) Distribution of G4-prone motifs according to the G4Hunter score. 1.2 means any sequence with a score between 1.2 and 1.399; 1.4 between 1.4 and 1.599, etc. These numbers are normalized by the total number of PQS found in bacteria, archaea, and compared with *Homo sapiens*. The first category represents 97.9% and 97.2% of all PQS sequences in bacteria and archaea, respectively. Note the log scale on the Y-axis.

Table 3. Genomic sequences sizes, GC%, total count of PQS, and mean frequencies of quadruplex motifs. Seq (total number of sequences), Median (median length of sequences), Short. (shortest sequence), Long. (longest sequence), GC % (average GC content), PQS (total number of predicted PQS), Mean f (mean frequency of predicted PQS per 1000 bp), Min f (lowest frequency of predicted PQS per 1000 bp), Max f (highest frequency of predicted PQS per 1000 bp). %PQS corresponds to the probability that any given nucleotide in the group or subgroup belongs to a G4-prone region (G4H > 1.2). Colors correspond to phylogenetic tree depiction.

Kingdom	Seq.	Median	Short	Long	GC %	PQS	Mean f	Min f	Max f	% PQS
Archeae	3387	1,686,930	100,212	13,399,915	46.51	7,927,775	1.21	0.04	15.31	3.58
Superphylum	Seq.	Median	Short	Long	GC %	PQS	Mean f	Min f	Max f	% PQS
BAT	320	1,180,629	164,795	3,506,105	43.07	421,678	1.16	0.05	8.42	3.49
Cren	379	1,808,184	210,860	6,451,204	43.05	1,009,660	1.56	0.09	9.44	4.75
Asgard	71	2,322,715	291,515	5,684,038	38.75	74,647	0.47	0.12	1.50	1.39
DPANN	309	832,169	100,212	6,604,953	39.22	219,058	0.70	0.08	4.20	2.18
Eury	2308	1,826,841	137,797	13,399,915	48.77	6,202,732	1.25	0.04	15.31	3.68
Phylum	Seq.	Median	Short	Long	GC %	PQS	Mean f	Min f	Max f	% PQS
Bathyarchaeota	128	1,208,976.5	200,493	3,506,105	46.29	245,162	1.54	0.23	8.42	3.00
Thaumarchaeota	192	1,173,909.5	164,795	3,441,569	40.93	176,516	0.91	0.05	5.32	2.73
Thermoproteales	147	1,581,744	242,587	3,969,448	45.86	513,053	2.07	0.11	7.38	6.31
Sulfolobales	118	2,223,757.5	210,860	3,034,024	38.20	200,842	0.79	0.34	4.58	2.38
Desulfurococcales	29	1,580,347	807,477	2,148,448	46.99	99,211	2.29	0.40	6.37	6.95
Verstraetearchaeota	18	1,171,913.5	419,172	1,937,662	46.76	40,586	1.83	0.10	3.43	5.50

Table 3. Cont.

Marsarchaeota	15	1,915,630	351,358	3,731,392	46.72	52,853	1.64	0.47	2.94	5.01
Geothermarchaeota	6	1,183,145.5	803,797	1,671,866	42.72	16,582	2.15	0.96	7.03	6.65
Nezhaarchaeota	2	1,332,140.5	1,315,707	1,348,574	43.53	2016	0.76	0.75	0.77	2.27
Korarchaeota	18	1,542,873	834,209	2,942,065	48.39	68,434	2.63	1.05	9.44	7.95
Unclassified Crenarchaeota	27	1,203,892	301,027	6,451,204	37.01	19,361	0.44	0.09	1.49	1.29
Lokiarchaeota	29	1,892,624	320,847	5,143,417	32.77	25,479	0.41	0.21	1.50	1.24
Odinarchaeota	1	1,460,710	1,460,710	1,460,710	38.05	1038	0.71	0.71	0.71	2.16
Thorarchaeota	29	2,770,204	291,515	4,389,059	46.55	40,006	0.60	0.24	1.18	1.76
Heimdallarchaeota	12	2,167,091	432,340	5,684,038	34.42	8124	0.27	0.12	0.50	0.82
Aenigmarchaeota	35	751,672	248,182	1,410,470	39.33	17,990	0.71	0.11	3.78	2.12
Nanohaloarchaeota	17	815,638	565,289	1,480,846	44.53	8672	0.48	0.09	1.82	1.50
Woesearchaeota	72	966,794.5	518,295	2,944,567	40.77	57,833	0.66	0.08	3.92	1.96
Pacearchaeota	60	719,507	279,432	6,604,953	33.74	37,675	0.56	0.08	2.99	1.73
Nanoarchaeota	25	577,110	204,081	1,162,239	32.83	9940	0.59	0.13	4.20	1.70
Micrarchaeota	39	887,931	658,716	1,333,875	50.41	42,298	1.17	0.15	2.86	3.47
Diapherotrites	19	568,419	302,064	1,130,899	37.42	6077	0.49	0.11	2.33	1.46
Unclassified DPANN	40	858,043.5	100,212	3,188,023	35.57	33,846	0.67	0.15	2.39	2.04
Hadesarchaeota	12	857,575	451,393	1,241,441	53.77	56,369	4.61	1.26	15.31	14.55
Persephonarchaeota	33	637,942	137,797	1,412,535	44.06	34,905	1.49	0.59	2.36	4.49
Thermococcales	60	1,867,904.5	207,909	2,388,527	46.77	191,492	1.72	0.47	7.53	5.15
Theinoarchaeota	2	4,165,806	3,559,548	4,772,064	41.57	5480	0.66	0.65	0.67	1.94
Methanofastidiosia	96	992,372	156,656	13,399,915	40.71	141,192	0.83	0.08	3.64	2.54
Methanococcales	24	1,717,483	1,207,361	1,936,387	32.01	15,065	0.39	0.20	0.86	1.19
Methanobacteriales	224	2,001,036	1,157,521	3,466,370	33.62	175,191	0.39	0.04	2.32	1.14
Methanopyrales	3	1,430,309	1,421,621	1,694,969	58.94	10,798	2.34	1.97	3.00	6.84
Methanomassilicoccales	91	1,404,109	640,223	2,641,216	56.22	257,340	1.85	0.22	4.41	5.38
Thermoplasmatales	135	1,621,237	593,453	2,816,557	42.71	246,832	1.13	0.11	7.03	3.42
Acidoprofundum/DHV2-2	11	1,731,076	519,420	2,981,805	40.55	16,609	1.21	0.29	4.12	3.59
Archaeoglobales	53	1,901,943	478,535	3,408,041	42.98	117,470	1.22	0.57	3.29	3.66
Methanosarcinales	279	2,913,215	208,261	5,751,492	44.99	845,394	1.19	0.15	7.52	3.54
Methanomicrobiales	146	2,228,967.5	622,799	3,978,804	54.97	783,172	2.38	0.23	7.20	7.07
Methanocellales	5	2,957,635	1,465,272	3,243,770	50.96	16,825	1.21	0.41	1.88	3.51
Halobacteriales	440	3,585,981	397,623	5,605,381	63.95	2,271,600	1.56	0.08	4.25	4.50
Unclassified Diaforarchaea	97	1,460,542	233,168	2,294,894	47.38	136,115	1.03	0.18	2.55	3.02
Unclassified other	597	1,400,198	258,312	7,416,915	46.88	862,962	1.02	0.07	5.16	3.00

The comparison of G4 prone sequences found in archaea with bacteria genomes revealed that in both domains, frequencies sharply decreased with G4HS as compared to the human genome, in which highly stable G4s are relatively more frequent (see Figure 3B). This result indicates an overall stronger relative selection pressure against stable G4 motifs in both archaea and bacteria as compared to humans, and likely most eukaryotes, as the relative number of G4Hunter high scoring motifs is even higher in yeast [63]. Guo and Bartel suggested that eukaryotes have robust machinery that globally unfolds RNA G-quadruplexes, whereas some bacteria have instead undergone evolutionary depletion of G-quadruplex-forming sequences [64]. Our analysis suggests that archaea behave like bacteria, except for the slight difference found for the most stable motifs (G4HS >2), which were less selected against in archaea than in bacteria.

3.2. Variation in Frequency for G4 Forming Sequences in Archaea

The total number of analyzed sequences in particular phylogenetic categories, together with a median length of the genome, shortest genome, longest genome, mean, minimal, and maximal observed frequency PQS per kbp, and total PQS counts are shown in Table 3. For this analysis, Archaea have been divided into five superphyla that form monophyletic assemblages (clades) in the most recent phylogenetic analysis and 41 subgroups that correspond to different taxonomic ranks (suffix *aeota* for phylum, candidate phylum, suffix *ales* for orders). Seven subgroups have an average GC content above 50%, the highest GC content being observed in *Halobacteriales* (63.95%), which is also the archaeal group containing the highest number of available genome sequences (440), all other groups have average GC contents below 50%.

The mean frequency of PQS per kbp for all archaeal genomes was 1.207. The lowest mean frequency was for the *Heimdallarchaeota* (0.273), followed by *Methanococcales* and *Methanobacteriales* (0.39). The highest density of PQS was found in the *Hadesarchaea* subgroup (4.607), followed by *Korarchaeota* (2.626). The highest absolute frequency of PQS was found in *Hadesarchaea* archaeon isolate WYZ-LMO6 with 15.3 PQS per 1000bp (i.e., one quadruplex every 65 bp), and the lowest frequency was found in *Methanobrevibacter* sp. 87.7: Interestingly, only 71 PQS were found in its 1.92 Mb long genome (Supplementary Table S2A). Detailed statistical characteristics for PQS frequencies per kbp (including mean, variance, outliers) are depicted in boxplots for all inspected subgroups (Figure 4). The *Hadesarchaea* subgroup has a higher PQS frequency in comparison to other subgroups. The comparison of the five main superphyla BAT, Cren, Asgard, Eury, and DPANN (*Diapherotrites*, *Parvarchaeota*, *Aenigmarchaeota*, *Nanoarchaeota*, and *Nanohaloarchaeota*) (Figure 1) revealed the highest mean PQS frequency in Cren superphylum (1.15) and the lowest in Asgard superphylum (0.48). However, the *Hadesarchaea* subgroup, which exhibits the highest frequency among subgroups, is found in the Eury superphylum. The detailed data for superphyla are in Supplementary Table S2B, for subgroups in Supplementary Table S2C.

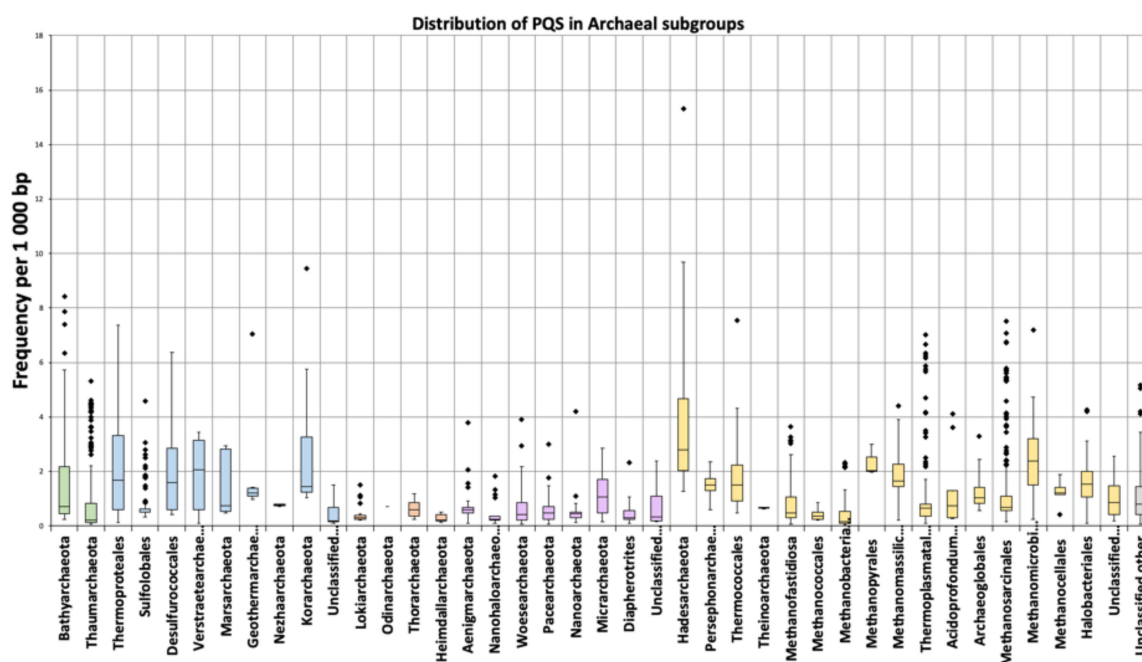


Figure 4. Frequencies of PQS in subgroups of analyzed archaeal genomes. Data within boxes span the interquartile range, and whiskers show the lowest and highest values within 1.5 interquartile range. Black points denote outliers. Horizontal black lines inside boxplots are median values.

A cluster dendrogram shows the similarities among subgroups based on the PQS data (Figure 5). This dendrogram shows that the *Hadesarchaeota* subgroup is the most distant one (the shortest branch length) compare to other subgroups. The cluster dendrogram based on PQS characteristics is similar to the phylogenetic relationships (see Figure 1). For example, all of the Asgard subgroups (*Odinarchaeota*, *Heimdallarchaeota*, *Thorarchaeota*, and *Lokiarchaeota*) lie close together, in one bigger cluster (Figure 5, left part). Other examples are the *Woesearchaeota*, *Aenigmarchaeota*, and *Nanoarchaeota* subgroups, which are members of the DPANN superphylum, and lie adjacent to each other in PQS based cluster tree. On the other hand, all of the subgroups with the prefix “-thermo”, indicative of high-temperature environments, are clustered together (*Thermoplasmatales*, *Thermococcales*, *Thermoproteales*, and *Geothermarchaeota*). These subgroups are relatively PQS rich, but lack phylogenetical proximity, suggesting that PQS richness does not rely on evolutionary proximity.

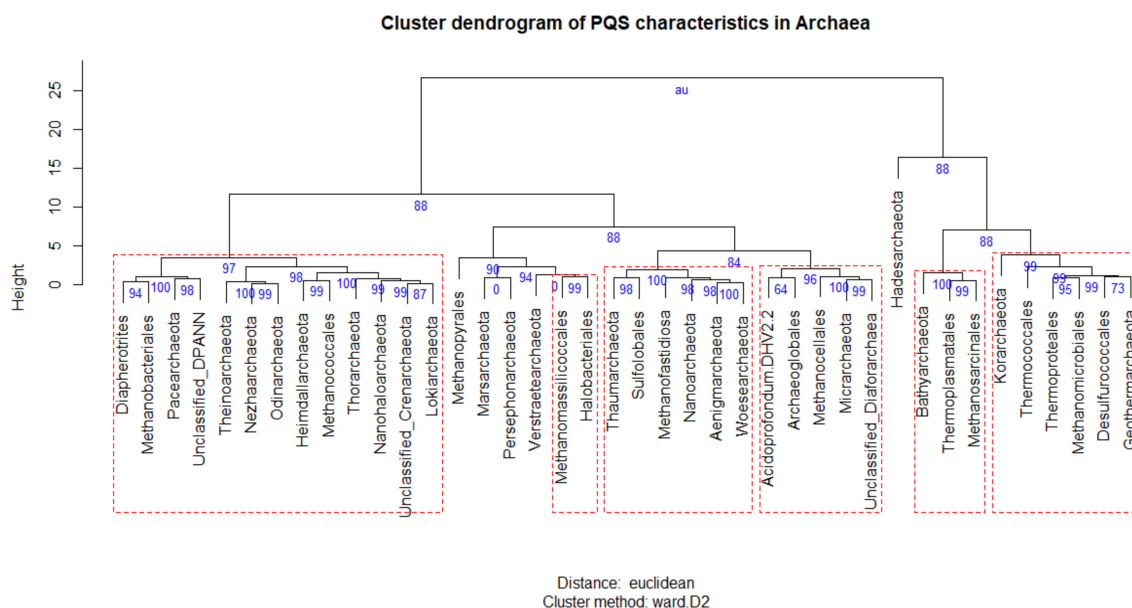


Figure 5. Cluster dendrogram of PQS characteristics of archaeal subgroups. Cluster dendrogram of PQS characteristics (Supplementary Table S4) was made in R v. 3.6.3 (code provided in Supplementary Table S4) using pvclust package with these parameters: Cluster method ‘ward.D2’, distance ‘euclidean’, number of bootstrap resamplings was 10,000. AU values are in blue and indicate the statistical significance of particular branching (values above 95 are equivalent to p -values lesser than 0.05). Statistically significant clusters are highlighted by red dashed rectangles.

We then analyzed the relationship between overall % GC content and PQS frequency (Figure 6). PQS frequencies tend to correlate with GC content as G4-prone motifs need to be relatively G-rich; however, there are interesting exceptions to this rule, and this correlation is poorer than anticipated. Ding et al. already noticed that *Methanomicrobia* and *Thermococci* have greater densities of PQS than the theoretical values based on the GC % of their genomes [35]. Organisms with higher than expected PQS frequencies based on their GC content (over 50% of the maximal observed PQS frequency, Figure 6) are highlighted in color; the whole figure is separated into smaller segments according to inspected G4Hunter score intervals. The most extreme outlier is *Hadesarchaea* archaeon, for which 51% of its genome has a G4Hunter score above 1.2, despite a GC content of 54%, i.e., only modestly above the 46.5% average for all sequences tested here, and far below the most GC rich archaea genomes. Cherry-picked examples of G-rich motifs with high G4 Hunter scores (G4HS) in *Hadesarchaea* archaeon are provided in Table 4. We have also carried out additional statistical evaluation of PQS differences between all groups and subgroups; detailed results are found in Supplementary Table S5. Nearly all comparisons were significant, i.e., there are significant differences between PQS frequencies of particular groups and subgroups.

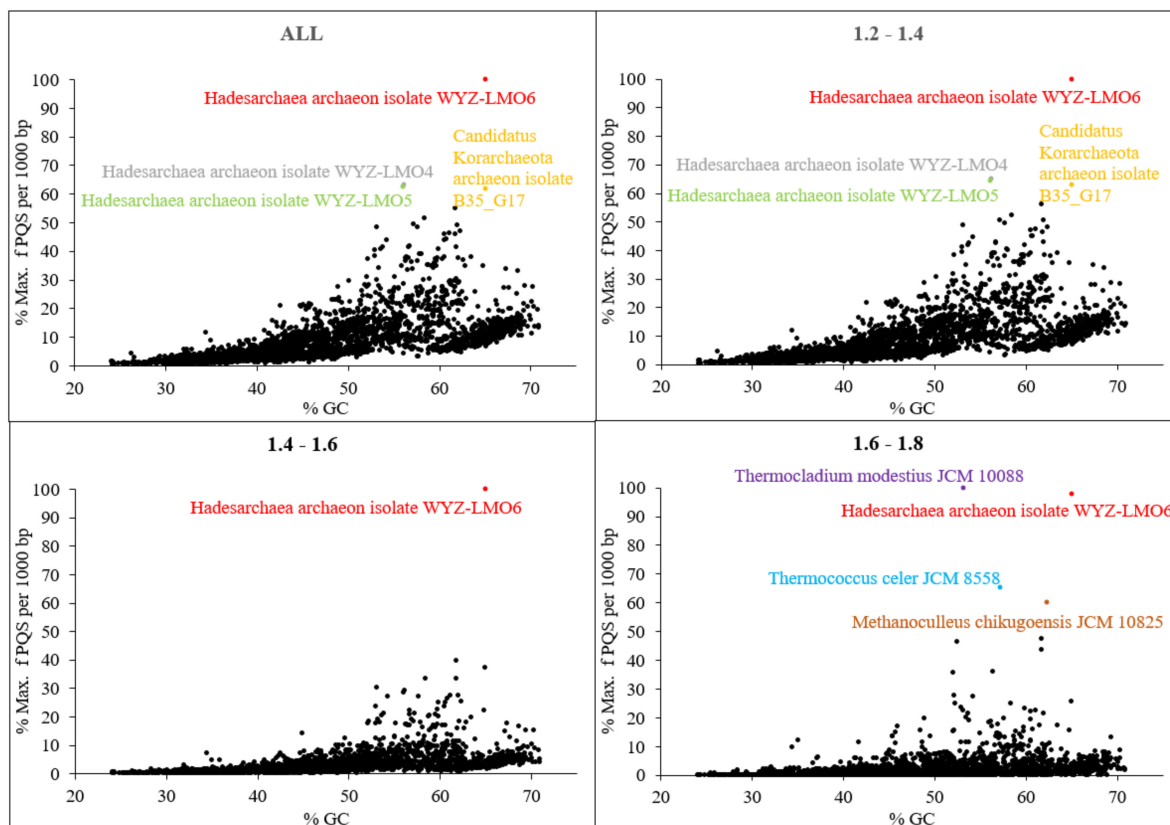


Figure 6. Relationship between the observed frequency of PQS per 1000 bp and GC content. Different G4Hunter score intervals are considered. In each G4Hunter score interval miniplot, frequencies were normalized according to the highest observed frequency of PQS. Organisms with max. frequency per 1000 bp greater than 50% are described and highlighted in color.

Table 4. Long G4-prone motifs with high G4HS found in *Hadesarchaea archeon*.

Name	Sequences (5' to 3')	G4 Hunter Score	IDS	CD
038_K	AGGCTGGGGGTGAGGGCGGTGGTGGGAAGGGAGGGGTGGGGGAGAAAACGAAGGGGGT	2.07	G4	Parallel
086_K	TGGGGAGGAGGGGAGGGAGGTGGGCTGGGGGGGCT	2.57	G4	Parallel
174_K	AGGGTGAAGGAGGTGCTGGGGGAAGGGAGGTGGGGAGGGGAGGTGGAGGGCTGGTGAAGGA	2.07	G4	Parallel
175_K	AGGGGAGGAGGTGGCCGTGGTGGGGCGGGGGAGGGGCGGGGTGGGGGGCCTGGGGGA	2.54	G4	Parallel
176_K	AGGAGGAGGTGAGGGACAGGGAGGAGGGAGGGGAGGGGGAAGGAGGAGGAGGAGGAGGGA	1.93	G4	Parallel
178_K	TGGTGGGGCGGGGGAGGGCGGGGTGGGGGGCCTGGGGGA	2.89	G4	Parallel
195_K	AGGGGAGGAGGTGGCCGTGGTGGGGCGGGGGAGGGGCGGGGTGGCCTCACGGA	1.91	G4	Parallel
196_K	AGGGGAGGAGGAGGGAGGGGGGAAGGAGGAGGAGGAGGAGGGA	2.22	G4	Parallel
245_K	GGGGTCGTGGGGGGGAGAGCTGGGGAGGAGGGAGGGAGGTGGGCTGGGGGGGCTGGGGAGGAGGAGGTGAGGGG	2.33	G4	Parallel
640_K	AGGGAGTGGGGGAGGGGAGGTGGAGGGGCT	2.38	G4	Parallel
642_K	TGGTGGGGCGGGGGAGGGGCGGGGT	2.93	G4	Hybrid*
643_K	AGGCTGGGGGTGAGGGCGGTGGTGGGAAGGGAGGGGTGGGGGAGAAAACGAAGGGGGT	2.07	G4	Parallel
644_K	AGGGCGGTGGTGGGAAGGGAGGGGTGGGGGA	2.41	G4	Parallel
645_K	GGCGGGGGGAGTCTTCATCTGGGGTAGGGG	1.74	G4	Parallel

* Sequence 642_K adopts a hybrid structure at room temperature, which is converted to a parallel conformation at high temperatures.

Figure 7 shows the relationship between GC percentage and mean PQS frequencies (or mean percentage of PQS length of the genome) in particular archaeal subgroups. Overall, we found some correlation (although far from perfect, as shown by $R^2 = 0.7$) between mean PQS frequencies (expressed as the mean fraction of nucleotides of the genomes involved a PQS motif) and increasing GC % content.

The highest mean percentage of PQS length of the genomes was found in subgroup *Hadesarchaea*, in which more than 10% of their genomes are involved in a potential PQS.

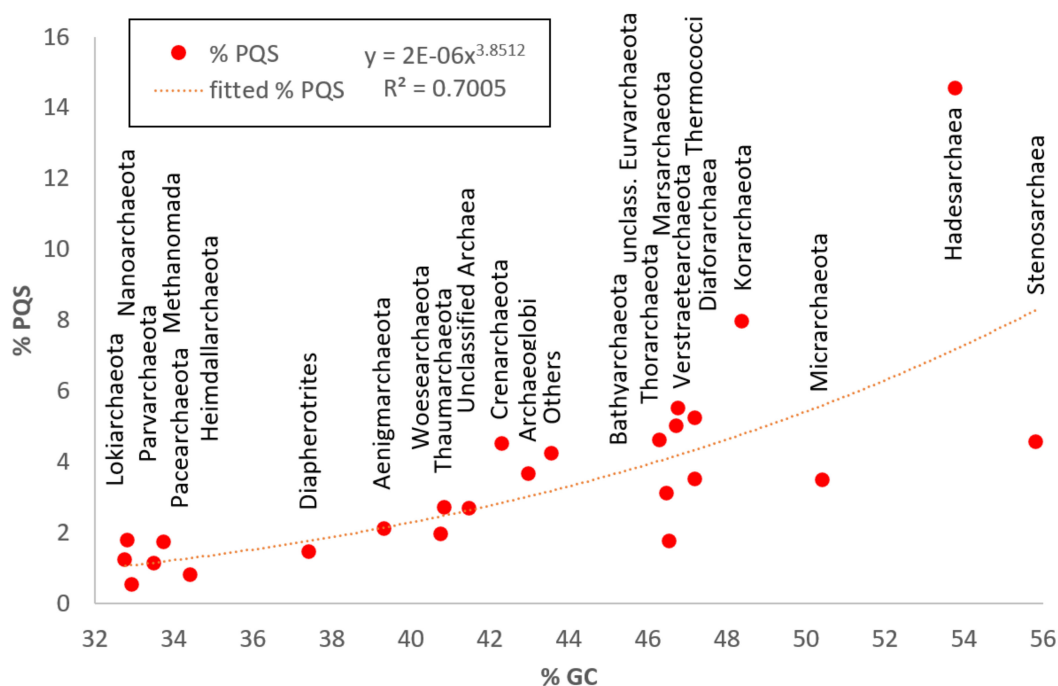


Figure 7. Relationship between GC percentage and % of PQS in genomes of particular archaeal subgroups. The Fitted equation with the R^2 coefficient is depicted on the top side of the plot.

3.3. Localization of PQS in Genomes

To evaluate the position of PQS in archaeal genomes, we downloaded the described “features” of all archaeal genomes and analyzed the presence of all PQS in annotated sequences (Figure 8). Overall, we find a higher density of G4-prone motifs in non-protein coding RNAs (tRNA, rRNA, and other ncRNA) than in protein-coding genes. G4 density in ncRNA is clearly above average genomic G4 density, while mRNA G4 density is close to the genomic average. This may derive in part from the observation that rRNA and tRNA genes are especially GC-rich in hyperthermophilic archaea, in order to stabilize folding under harsh conditions [65]. On the other hand, we can probably expect a stronger selection pressure against the formation of intramolecular quadruplexes within the relatively small tRNA core, as this would disrupt its three-dimensional shape and alter its biological function. In line with this hypothesis, the PQS frequencies are actually lower in tRNA than in ncRNA and rRNA [66]. Interestingly, the 5' end of some human tRNA genes is often G-rich and has been reported to allow G4 formation: Ivanov and colleagues have shown that mature cytoplasmic tRNAs are cleaved during stress response to produce tRNA fragments that function to repress translation in vivo and that these bioactive tRNA fragments assemble into intermolecular RNA G4s [67]. The 5' fragment of tRNA^{Ala} involves a predominant hairpin structure that starts with the 5'-GGGGGU motif, allowing the formation of tetramolecular quadruplex structures with five tetrad layers. Interestingly, tRNA-derived fragments have also been described in archaea. For example, a 26-residue-long fragment (5' GGGUUGGUGGUCUAGUCUGGUUAUGA) originating from the 5' part of valine tRNA is the most abundant tRNA fragment in *Haloferox volcanii* [68]. This fragment, while exhibiting a relatively G-rich 5' end (starting with GGGUUGG), may, in principle, allow intermolecular quadruplex formation as well.

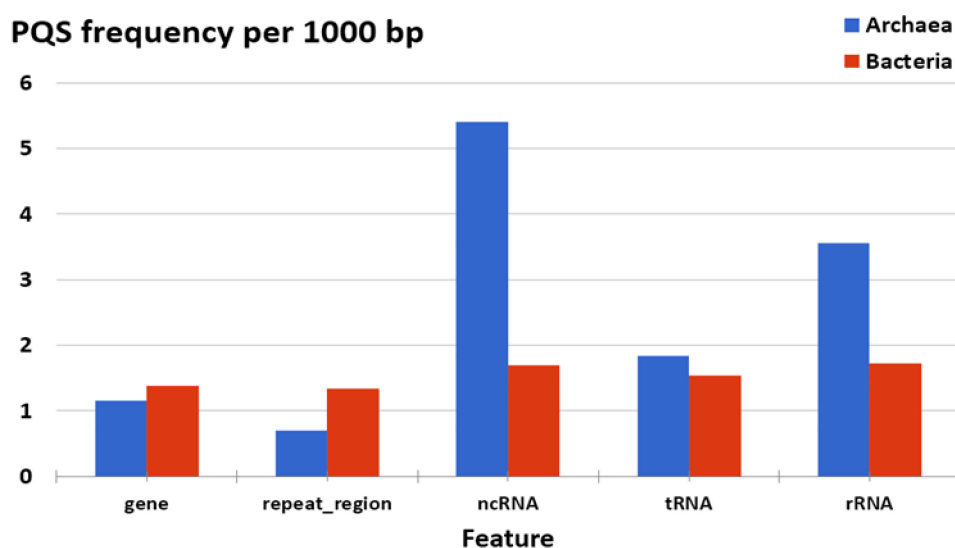


Figure 8. Differences in PQS frequency by DNA locus. The chart shows PQS frequencies normalized per 1000 bp annotated locations from the NCBI database and shows a comparison between Archaea and Bacteria. Archaea G4-prone motifs are strongly over-represented in ncRNA and rRNA compared to the average G4 density in Archaea (mean $f = 1.207$), but also compared to bacteria. PQS count is provided in Supplementary Table S3 Excel file.

Unfortunately, other features in archaeal genomes are so poorly annotated that we cannot use these data for evaluation. Comparison of PQS frequencies in annotated sequences with analyses of Bacteria shows the same trend for ncRNA, rRNA, protein-coding gene, and tRNA features. In contrast, the frequency in bacteria for ncRNA is 1.7 per kbp, and the frequency in archaea for ncRNA is 5.3 per kbp. On the other hand, the PQS frequency in repeat regions is lower in archaea than in the bacteria genome. We have to take into account that the data could be influenced by poor annotation in archaea genomes, and also by a low number of annotated sequences in Archaea; only 141 representative archaeal genomes are annotated, compared to 1627 representative bacteria annotated genomes. The strong abundance of the PQS in ncRNA compare to other locations pointing to its functional relevance. ncRNAs are present in the cells as single-stranded molecules in contrast to DNA, and therefore, they can easily adopt the G4 structures as a part of their 3D arrangement similarly to mRNAs [69,70]. It has been shown that ncRNAs play important roles in many cellular processes, including the regulation of gene transcription, post-transcriptional, and epigenetic regulations [71,72].

Other specific regions, such as replication origins or promoter regions, were not included in this graph. The oriC 10.0 database (<http://tubic.org/doric/public/index.php>) contains 226 archaeal origins of replication obtained by both in vitro studies and in silico predictions ([73]), prediction and experimental data are available for the *Thermococcales* [74,75], the *Haloarchaea*, and the *Sulfolobales* [76]. Archaeal replicators, as in bacteria, are composed of three main elements: A cluster of binding sites for the initiator Cdc6, the DNA unwinding element (DUE), and binding sites for regulatory proteins [75]. Interestingly, it was found in several *Haloarchaea* species that a specific (TGGGGGGG) motif occurs in one of the two origins of replication (oriC1) [77]. This long G-rich motif was shown to be necessary for efficient replication initiation in *Haloarcula hispanica* [78,79] and predicted to be prone to inter-molecular quadruplex formation.

3.4. Experimental Demonstration of Quadruplex Formation In Vitro

Next, we selected a few DNA G4-prone motifs found in *Hadesarchaea* and experimentally tested if they formed a G4 structure under classical conditions. As inferred from isothermal difference spectra (IDS) (Figure 9a) and circular dichroism (CD) spectra (Figure 9b), all motifs clearly formed G-quadruplexes at room temperature. However, as these motifs are found in an archeon expected to live

at a high temperature, we also recorded the spectra at 80 °C. As shown in Figure 9c, these quadruplexes were thermally stable and still formed at high temperatures. Of note, most spectra are indicative of a parallel fold. This bias is the result of a high threshold for G4Hunter (all motifs have scores > 1.7). As a consequence, these motifs are very G-rich, with runs of G separated by short spacers, often 1–2 nt. As short loops tend to be propeller-type, this sequence bias will favor a parallel conformation.

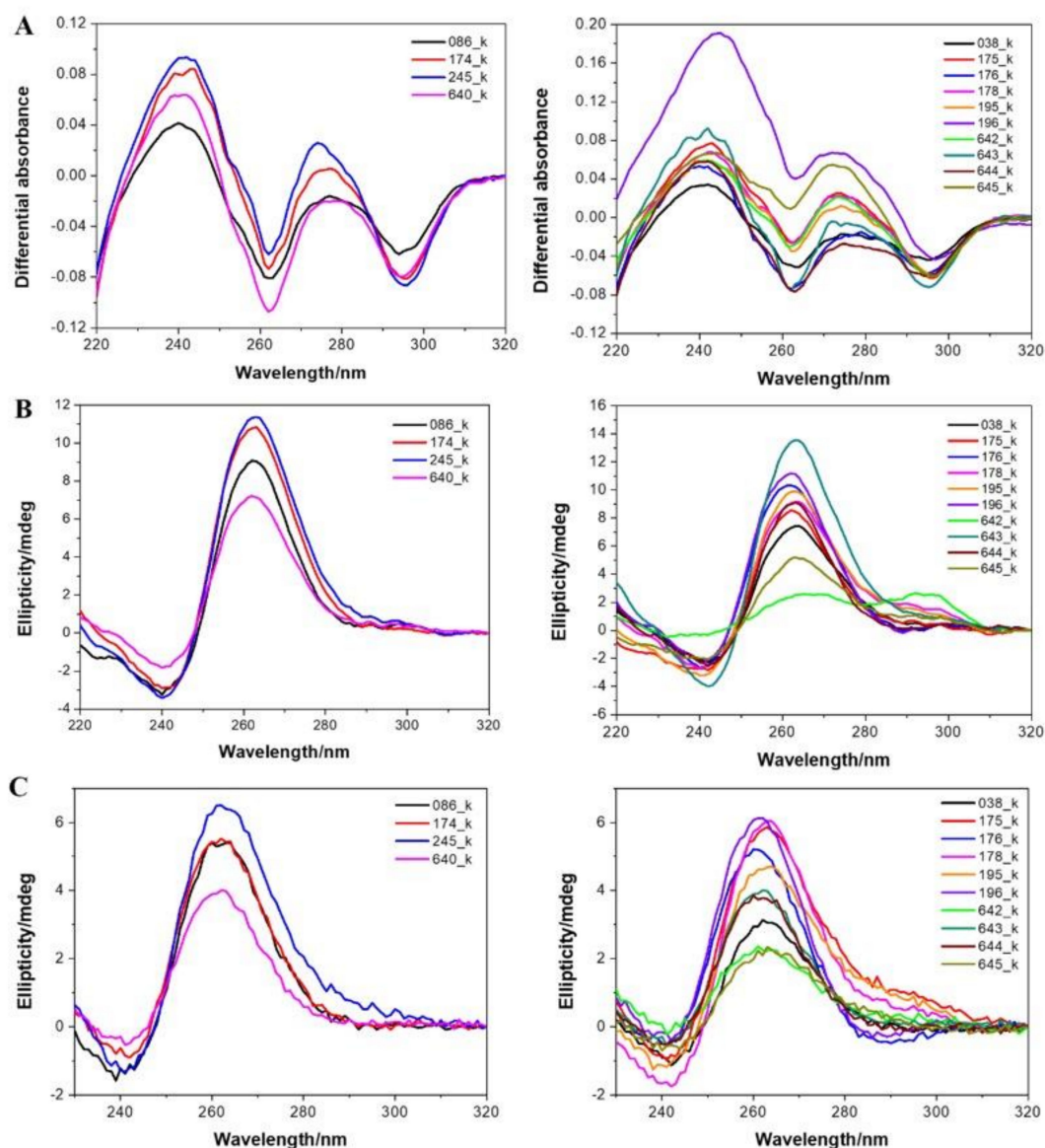


Figure 9. Experimental evidence for quadruplex formation with archaea sequences. Isothermal differential absorbance (IDA; panel A) and circular dichroism (CD; panels B and C) spectra of *Hadesarchaea archeon* DNA sequences were recorded at 20 °C (panels A and B) or at a high temperature (80 °C) for CD (panel C).

3.5. G4-Binding Proteins from Archaea

Given that G4-prone motifs are found in Archaea, and actually extremely abundant in some subgroups, it was interesting to check if potential helicases are present to solve these structures. A number of DNA and RNA G4-helicases have been identified in eukaryotes, e.g., Pif1, DOG, Rhau/DHX36, WRN, BLM; for a review [80]. Little or no experimental data is currently available on archaea enzymes able to unfold G-quadruplexes. As RecQ has been reported to unfold G4 structures

in bacteria, we searched for RecQ homologs in Archaea. A BLASTp search using RecQ (UNIPROT ID: P15043) from *E. coli* as a query revealed 1206 homologous protein sequences in a archaeal domain with an E-value cut-off = 0.0001. A listing of all candidates identified is presented in supplementary information (Supplementary Table S6). Five proteins have an identity with G-quadruplex RecQ resolvase higher than 50%, and 312 proteins have more than 50% aa positives hits in the sequence, suggesting that they share the G4 unfolding functionality in archaea genomes. Besides protein actively unfolding G4 structures, other peptides may actually bind to single-strand G-rich sequences and passively contribute to G4 unfolding by conformational selection. This is the case for a single-strand binding protein isolated from *Methanococcus jannaschii*, which was used to design an assay to detect G4 formation [79]. Apart from proteins that actively or passively unfold quadruplexes, others may bind to and sometimes promote G4 formation. The amino acid composition of 77 G-quadruplex binding proteins from *Homo sapiens* revealed unique features of quadruplex binding proteins, with prominent enrichment for glycine (G) and arginine I [31]. Human-binding proteins share a 20 amino acid long motif/domain (RGRGR GRGGG SGGSG GRGRG), which is similar to the previously described RG-rich domain of the FMR1 G-quadruplex binding protein. The search for this 20 amino acid-long motif in archaea proteome found 23 hits/potential G-quadruplex binding proteins with an E-value threshold of 0.05; the identity was found, e.g., for RNA DEAD box helicase or for two 30S ribosomal proteins S4 (Supplementary Table S6, list 2). We searched protein sequences in the proteome of the mesophilic archaeon *Methanosarcina mazei* (for which the largest amount of proteins is known) for the presence of this motif. For highly significant p values ($p < 10^{-6}$), we found four proteins with a potential quadruplex-binding motif (Supplementary Table S7), while significantly more (193) hits were found for p-values $< 1 \times 10^{-5}$. Three of them are without any known function (DUF134 domain-containing protein, PGF-pre-PGF domain-containing protein, and DUF5320 domain-containing protein). Even if the full proteome of *Hadesarchaea archaeon* is not known, it is interesting to note that this RG-domain is present in a number of putative proteins. In addition, while a true RecQ homolog was not found, one *Hadesarchaea archaeon* 600aa-polypeptide has a good similarity with RecQ in its N-terminal half (Supplementary Table S8). The presence of the NIQI motif in the “DNA-directed RNA polymerase subunit” is also interesting and possibly logical, given the necessity of unraveling G-quadruplexes during transcription. The presence in archaeal genomes of potential G4-binding and G4-unfolding proteins supports the formation of quadruplex structures in archaeal cells.

4. Discussion

We provide here the first comprehensive study of PQS occurrences, frequencies, and distributions in archaeal genomes. The overall analysis made on global frequency hides extreme differences between species and subgroups, which can be explained by differences in GC content and possibly codon usage.

At one end of the G4 spectrum, some subgroups of archaea, such as *Parvoarchaeota* or *Heimdallarchaeota*, have very low PQS frequencies, and PQS cover 1% or less of their genomes. In sharp contrast, we found an unprecedented enrichment of PQS for some subgroups, often living under extreme conditions. For example, over 50% of the genome of *Hadesarchaea archaeon* may potentially adopt a quadruplex fold. This *Hadesarchaea* is living under extreme conditions, as it was found in South African gold mines 3 km underground, without light and oxygen (*Hades* is the Greek god of the underworld). Following this analysis, we used the BioSample NCBI database [78] to compare the living environment of the archaea organisms with the highest PQS frequencies. Data for all genomes with PQS frequency above 6 per kbp are shown in Table 5. A majority of organisms with extremely high PQS frequencies are found in hot springs sediments or in deep-sea hydrothermal vent sediments, and this high PQS frequency may be associated with their extremophilic life, although more work will be necessary to compare G4 density in acidophilic, thermophilic, halophilic and psychrophilic organisms. For example, in bacteria, in the Gram-positive subgroup *Deinococcus-Thermus*, a high PQS frequency was associated with their extremophilic origin [35,81], while the gram-negative extremophilic bacteria subgroup *Thermotogae* are among organisms with a low PQS frequency [33]. We suggest

that the high stability of G4 structures compare to dsDNA structure could play important roles in archaea and Gram-positive extremophiles organisms. We then experimentally confirmed G4 formation with a few archaea sequences to confirm that our in silico predictions are verified: All predicted experimentally tested formed stable G-quadruplexes in vitro. This absence of false positives is hardly surprising given that we chose high scoring motifs. From our published [41] and unpublished data on now over 500 sequences, false positives for sequences with scores above 1.5 are extremely rare (<1.5%), and we have yet to find a false positive with a score > 1.75. Some of the sequences considered were long and may even allow the formation of two juxtaposed G4 structures. In a few cases, we can even propose a topology, as for example, TGGTGGGGGCGGGGGAGGGGCGGGGGT (642K), in which the predicted guanine tracks (underlined) may either be: TGGTGGGGGCGGGGGAGGGGCGGGGGT or TGGTGGGGGCGGGGGAGGGGCGGGGGT, and different folds may result from these possibilities (the latter would be likely parallel, as experimentally observed at 80 °C, while the former may adopt a non-parallel fold, as observed at room temperature). Note, however, that G4 hunter does not make any hypothesis on the G tracts involved in G4 formation, in contrast with Quadparser, for example, where one actively seeks the four runs of G involved in G-quartet formation. G4 formation is (still) full of surprises, and correctly predicting which runs (or individual guanines) participate in G-quartet formation is far from trivial and requires extensive experimental validation.

The extreme enrichment found in some archaea challenges our existing views on “noncanonical” DNA structures to which G-quadruplexes belong, as it is plausible that a substantial part of the *Hadesarchaea* genome may be packed into G-quadruplex structures. The complementary C-rich strand may also fold into a different quadruplex structure called the i-motif [82] that is favored by acidic pH. Further studies will be dedicated to i-DNA formation in Archaea.

Table 5. Detailed characteristics of archaeal species with PQS frequency per 1000 bp greater than 6.00. Living environments data were obtained from the BioSample NCBI database [83].

Organism Name	GC Content	PQS f	% PQS	Living Environment (Isolated from)
<i>Hadesarchaea archaeon</i> isolate WYZ-LMO6	65.01	15.310	51.15	Hot springs sediment, Yellowstone NP, USA
<i>Hadesarchaea archaeon</i> isolate WYZ-LMO4	56.17	9.685	31.10	Hot springs sediment, Jinze hot spring, China
<i>Hadesarchaea archaeon</i> isolate WYZ-LMO5	56.04	9.581	30.69	Hot springs sediment, Jinze hot spring, China
<i>Korarchaeota archaeon</i> isolate B35_G17	65.01	9.445	28.80	Deep-sea hydrothermal vent sediments, Guaymas Basin, Gulf of California, Mexico
<i>Bathyarchaeota archaeon</i> B23	61.78	8.418	26.12	Deep-sea hydrothermal vent sediments, Guaymas Basin, Gulf of California, Mexico
<i>Bathyarchaeota archaeon</i> isolate M10_bin139	58.42	7.858	24.55	Deep-sea hydrothermal vent sediments, Guaymas Basin, Gulf of California, Mexico
<i>Thermococcus celer</i> JCM 8558	57.21	7.534	24.52	Solfataric marine water hole on a beach of Vulcano, Italy
<i>Methanosaeta harundinacea</i> isolate UBA152	62.01	7.518	23.12	Waste water, Suncor tailings pond 6, Canada
<i>Bathyarchaeota archaeon</i> isolate B23_G15	57.67	7.397	22.90	Deep-sea hydrothermal vent sediments, Guaymas Basin, Gulf of California, Mexico
<i>Thermocladium modestius</i> JCM 10088	53.14	7.381	25.59	Mud from a spring pool, Noji-onsen, Fukushima, Japan
<i>Methanoculleus chikugoensis</i> JCM 10825	62.36	7.198	22.90	Paddy field soil, Chikugo, Fukuoka, Japan

Table 5. Cont.

Organism Name	GC Content	PQS f	% PQS	Living Environment (Isolated from)
<i>Methanosaeta harundinacea</i> isolate UBA281	61.14	7.089	21.80	Wastewater, North Alberta, Canada
<i>Geothermarchaeota</i> archaeon ex4572_27	60.54	7.032	22.01	Deep-sea hydrothermal vent sediments, Guaymas Basin, Gulf of California, Mexico
<i>Thermoplasmata</i> archaeon isolate CSSed11_322R1	61.82	7.028	22.57	Hypersaline soda lake sediment, Kulunda Steppe, Russia
<i>Methanosarcinales</i> archaeon Methan_02	60.8	6.738	20.67	Anaerobic digester metagenome, Australia
<i>Methanosaeta harundinacea</i> 6Ac	60.6	6.721	20.66	isolated from an upflow anaerobic sludge blanket reactor treating beer-manufacture wastewater in Beijing, China. (ref PMID:16403877)
<i>Thermoplasmatales</i> archaeon ex4484_36	54.25	6.673	21.15	Deep-sea hydrothermal vent sediments, Guaymas Basin, Gulf of California, Mexico
<i>Aeropyrum camini</i> SY1 = JCM 12091	56.73	6.370	19.72	Deep-sea hydrothermal vent chimney, the Suiyo Seamount in the Izu-Bonin Arc, Japan
<i>Bathymarchaeota</i> archaeon isolate B46_G17	61.92	6.332	19.03	Deep-sea hydrothermal vent sediments, Guaymas Basin, Gulf of California, Mexico
<i>Thermoplasmata</i> archaeon isolate B14_G15	53.83	6.327	20.11	Deep-sea hydrothermal vent sediments, Guaymas Basin, Gulf of California, Mexico
<i>Thermoplasmata</i> archaeon isolate B23_G1	53.66	6.240	19.72	Deep-sea hydrothermal vent sediments, Guaymas Basin, Gulf of California, Mexico
<i>Pyrobaculum neutrophilum</i> V24Sta	59.91	6.233	19.52	isolated from a hot spring in Iceland
<i>Thermoplasmata</i> archaeon isolate B23_G9	52.98	6.164	19.65	Deep-sea hydrothermal vent sediments, Guaymas Basin, Gulf of California, Mexico

Hadesarchaea archaeon isolates WYZ-LMO4, WYZ-LMO5, WYZ-LMO6 are archaeal species isolated from hydrothermal spring sediments. Besides high temperatures, often above 50 °C, these ecological niches usually have high salinity. Interestingly, most G-quadruplexes withstand high temperatures (their melting point is often above 70 °C) and are further stabilized by positively charged ions such as K⁺ and Na⁺ [84,85]. Such conditions may have naturally favored G-quadruplexes over duplexes. It also highlights one of the consequences of a high GC %: G4-prone motifs become more frequent (Figure 5). In addition, all hyperthermophilic organism genomes encode a reverse gyrase, which positively supercoil DNA, possibly to protect the genome [86]. In future studies, it would be very interesting to carry out a genome-wide wet-lab experiment, for example, direct DNA sequencing of G-quadruplex loci as described in [87,88] or direct visualization of G-quadruplexes in living cells using specific antibodies, such as BG4 [89].

5. Conclusions

Overall, our results indicate that archaea are, like eukaryotes and bacteria, prone to G-quadruplex formation: G-quadruplexes are here, there, and everywhere! Important differences in G4 densities were found among species, and experimental validation was obtained in vitro for a few candidate sequences. Follow-up studies may check if specific archaeal PQS loci—for example, in important genes, show some phylogenetic conservation. If confirmed, this could serve as a new (additional) phylogenetic marker and give us some extended clues about the evolution and function of G-quadruplex forming sequences in Archaea. This study will stimulate further studies on G4 presence in Archaea, and help to establish whether some regulatory mechanisms may only apply to a given domain or be truly universal.

Supplementary Materials: The following are available online at <http://www.mdpi.com/2218-273X/10/9/1349/s1/>, Figure S1: Experimental evidence for G4 formation with Hadesarchaea sequences at high temperature; Table S1: The accession codes and phylogenetic classification of all archaeal genomic DNA sequences, Table S2: Overall results of PQS frequencies found in each analyzed genomic sequence (all (A), superphylum (B) or phylum (C)) together with GC content, sequence length and other parameters, Table S3: Feature counts, Table S4: PQS characteristics used for the dendrogram shown in Figure 6, Table S5: Statistical evaluation, Table S6: BLASTp search for RecQ and NIQI in Archaea, Table S7: FIMO search for putative quadruplex binding motif, Table S8: The most similar protein of RecQ (E. coli) in Hadesarchaea archaeon.

Author Contributions: Conceptualization, V.B. and J.-L.M.; methodology, P.K.; software, O.P., J.Š., and P.P.; validation, V.B., P.K. and M.B.; formal analysis, M.B.; resources, M.B., P.F., V.D.C.; data curation, V.B., M.B., P.F., V.D.C., J.-L.M.; Experimental validation, Y.L., D.V.; writing—original draft preparation, V.B., M.F. and J.-L.M.; writing—review and editing, P.P. H.M., T.S.T., P.F., H.M., V.D.C., J.-L.M.; visualization, V.B., J.-L.M.; supervision, V.B., J.-L.M.; project administration, V.B., M.F.; funding acquisition, V.B., M.F., J.-L.M. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the Czech Science Foundation (18-15548S) and by the SYMBIT project Reg. no. CZ.02.1.01/0.0/0.0/15_003/0000477 financed from the ERDF.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Woese, C.R.; Fox, G.E. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proc. Natl. Sci. Acad. USA* **1977**, *74*, 5088–5090. [[CrossRef](#)]
2. Olsen, G.J.; Woese, C.R. Archaeal genomics: An overview. *Cell* **1997**, *89*, 991–994. [[CrossRef](#)]
3. Forterre, P. Archaea: What can we learn from their sequences? *Curr. Opin. Genet. Dev.* **1997**, *7*, 764–770. [[CrossRef](#)]
4. Grüber, G.; Manimekalai, M.S.S.; Mayer, F.; Müller, V. ATP synthases from archaea: The beauty of a molecular motor. *Biochim. Biophys. Acta* **2014**, *1837*, 940–952. [[CrossRef](#)]
5. Bolhuis, A. The archaeal Sec-dependent protein translocation pathway. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **2004**, *359*, 919–927. [[CrossRef](#)]
6. Samson, R.Y.; Dobro, M.J.; Jensen, G.J.; Bell, S.D. The Structure, Function and Roles of the Archaeal ESCRT Apparatus. *Subcell. Biochem.* **2017**, *84*, 357–377. [[CrossRef](#)]
7. Spang, A.; Eme, L.; Saw, J.H.; Caceres, E.F.; Zaremba-Niedzwiedzka, K.; Lombard, J.; Guy, L.; Ettema, T.J.G. Asgard archaea are the closest prokaryotic relatives of eukaryotes. *PLoS Genet.* **2018**, *14*, e1007080. [[CrossRef](#)] [[PubMed](#)]
8. Da Cunha, V.; Gaia, M.; Nasir, A.; Forterre, P. Asgard archaea do not close the debate about the universal tree of life topology. *PLoS Genet.* **2018**, *14*, e1007215. [[CrossRef](#)]
9. Adam, P.S.; Borrel, G.; Brochier-Armanet, C.; Gribaldo, S. The growing tree of Archaea: New perspectives on their diversity, evolution and ecology. *ISME J.* **2017**, *11*, 2407. [[CrossRef](#)]
10. Spang, A.; Caceres, E.F.; Ettema, T.J.G. Genomic exploration of the diversity, ecology, and evolution of the archaeal domain of life. *Science* **2017**, *357*. [[CrossRef](#)]
11. Pennisi, E. Survey of archaea in the body reveals other microbial guests. *Science* **2017**, *358*, 983. [[CrossRef](#)] [[PubMed](#)]
12. Chaudhary, P.P.; Conway, P.L.; Schlundt, J. Methanogens in humans: Potentially beneficial or harmful for health. *Appl. Microbiol. Biotechnol.* **2018**, *102*, 3095–3104. [[CrossRef](#)]
13. Vuillemin, A.; Wankel, S.D.; Coskun, Ö.K.; Magritsch, T.; Vargas, S.; Estes, E.R.; Spivack, A.J.; Smith, D.C.; Pockalny, R.; Murray, R.W. Archaea dominate oxic seafloor communities over multimillion-year time scales. *Sci. Adv.* **2019**, *5*, eaaw4108. [[CrossRef](#)]
14. Jain, S.; Caforio, A.; Driessen, A.J.M. Biosynthesis of archaeal membrane ether lipids. *Front. Microbiol.* **2014**, *5*, 641. [[CrossRef](#)] [[PubMed](#)]
15. Nobu, M.K.; Narihiro, T.; Kuroda, K.; Mei, R.; Liu, W.-T. Chasing the elusive Euryarchaeota class WSA2: Genomes reveal a uniquely fastidious methyl-reducing methanogen. *ISME J.* **2016**, *10*, 2478–2487. [[CrossRef](#)] [[PubMed](#)]

16. Aouad, M.; Borrel, G.; Brochier-Armanet, C.; Gribaldo, S. Evolutionary placement of Methanonatronarchaea. *Nat. Microbiol.* **2019**, *4*, 558–559. [[CrossRef](#)]
17. Forterre, P. The universal tree of life: An update. *Front. Microbiol.* **2015**, *6*. [[CrossRef](#)]
18. Dombrowski, N.; Lee, J.-H.; Williams, T.A.; Offre, P.; Spang, A. Genomic diversity, lifestyles and evolutionary origins of DPANN archaea. *FEMS Microbiol. Lett.* **2019**, *366*, fnz008. [[CrossRef](#)]
19. Gaia, M.; Forterre, P. The Tree of Life. In *Molecular Mechanisms of Microbial Evolution (Grand Challenges in Biology and Biotechnology)*; Rampelotto, P.H., Ed.; Springer: New York, NY, USA, 2018.
20. Sun, Z.-Y.; Wang, X.-N.; Cheng, S.-Q.; Su, X.-X.; Ou, T.-M. Developing Novel G-Quadruplex Ligands: From Interaction with Nucleic Acids to Interfering with Nucleic Acid–Protein Interaction. *Molecules* **2019**, *24*, 396. [[CrossRef](#)]
21. Harkness, R.W.; Mittermaier, A.K. G-quadruplex dynamics. *BBA Proteins Proteomics* **2017**, *1865*, 1544–1554. [[CrossRef](#)]
22. Siddiqui-Jain, A.; Grand, C.L.; Bearss, D.J.; Hurley, L.H. Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress c-MYC transcription. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 11593–11598. [[CrossRef](#)] [[PubMed](#)]
23. Lee, S.C.; Zhang, J.; Strom, J.; Yang, D.; Dinh, T.N.; Kappeler, K.; Chen, Q.M. G-Quadruplex in the NRF2 mRNA 5' Untranslated Region Regulates De Novo NRF2 Protein Translation under Oxidative Stress. *Mol. Cell. Biol.* **2016**, *37*. [[CrossRef](#)] [[PubMed](#)]
24. Crenshaw, E.; Leung, B.P.; Kwok, C.K.; Sharoni, M.; Olson, K.; Sebastian, N.P.; Ansaloni, S.; Schweitzer-Stenner, R.; Akins, M.R.; Bevilacqua, P.C.; et al. Amyloid Precursor Protein Translation is Regulated by a 3'UTR Guanine Quadruplex. *PLoS ONE* **2015**, *10*. [[CrossRef](#)] [[PubMed](#)]
25. Gage, H.L.; Merrick, C.J. Conserved associations between G-quadruplex-forming DNA motifs and virulence gene families in malaria parasites. *BMC Genomics* **2020**, *21*, 236. [[CrossRef](#)]
26. Gazanion, E.; Lacroix, L.; Alberti, P.; Gurung, P.; Wein, S.; Cheng, M.; Mergny, J.; Gomes, A.; Lopez-Rubio, J. Genome wide distribution of G-quadruplexes and their impact on gene expression in malaria parasites. *PLoS Genetics* **2020**. [[CrossRef](#)]
27. Cahoon, L.A.; Seifert, H.S. An alternative DNA structure is necessary for pilin antigenic variation in *Neisseria gonorrhoeae*. *Science* **2009**, *325*, 764–767. [[CrossRef](#)]
28. Thakur, R.S.; Desingu, A.; Basavaraju, S.; Subramanya, S.; Rao, D.N.; Nagaraju, G. Mycobacterium tuberculosis DinG is a structure-specific helicase that unwinds G4 DNA implications for targeting g4 dna as a novel therapeutic approach. *J. Biol.* **2014**, *289*, 25112–25136.
29. Mishra, S.K.; Shankar, U.; Jain, N.; Sikri, K.; Tyagi, J.S.; Sharma, T.K.; Mergny, J.-L.; Kumar, A. Characterization of G-Quadruplex Motifs in espB, espK, and cyp51 Genes of Mycobacterium tuberculosis as Potential Drug Targets. *Mol. Ther. Nucleic Acids* **2019**, *16*, 698–706. [[CrossRef](#)]
30. Brazda, V.; Haronikova, L.; Liao, J.C.; Fojta, M. DNA and RNA Quadruplex-Binding Proteins. *Int. J. Mol. Sci.* **2014**, *15*, 17493–17517. [[CrossRef](#)]
31. Brázda, V.; Červeň, J.; Bartas, M.; Mikysková, N.; Coufal, J.; Pečinka, P. The Amino Acid Composition of Quadruplex Binding Proteins Reveals a Shared Motif and Predicts New Potential Quadruplex Interactors. *Molecules* **2018**, *23*, 2341. [[CrossRef](#)]
32. Ribeyre, C.; Lopes, J.; Boulé, J.-B.; Piazza, A.; Guédin, A.; Zakian, V.A.; Mergny, J.-L.; Nicolas, A. The yeast Pif1 helicase prevents genomic instability caused by G-quadruplex-forming CEB1 sequences in vivo. *PLoS Genet.* **2009**, *5*, e1000475. [[CrossRef](#)]
33. Bartas, M.; Čutová, M.; Brázda, V.; Kaura, P.; Šťastný, J.; Kolomazník, J.; Coufal, J.; Goswami, P.; Červeň, J.; Pečinka, P. The Presence and Localization of G-Quadruplex Forming Sequences in the Domain of Bacteria. *Molecules* **2019**, *24*, 1711. [[CrossRef](#)] [[PubMed](#)]
34. Marguet, E.; Forterre, P. DNA stability at temperatures typical for hyperthermophiles. *Nucleic Acids Res.* **1994**, *22*, 1681–1686. [[CrossRef](#)] [[PubMed](#)]
35. Ding, Y.; Fleming, A.M.; Burrows, C.J. Case studies on potential G-quadruplex-forming sequences from the bacterial orders Deinococcales and Thermales derived from a survey of published genomes. *Sci. Rep.* **2018**. [[CrossRef](#)]
36. Kota, S.; Dhamodharan, V.; Pradeepkumar, P.I.; Misra, H.S. G-quadruplex forming structural motifs in the genome of *Deinococcus radiodurans* and their regulatory roles in promoter functions. *Appl. Microbiol. Biotechnol.* **2015**, *99*, 9761–9769. [[CrossRef](#)] [[PubMed](#)]

37. Mishra, S.; Chaudhary, R.; Singh, S.; Kota, S.; Misra, H.S. Guanine Quadruplex DNA Regulates Gamma Radiation Response of Genome Functions in the Radioresistant Bacterium *Deinococcus radiodurans*. *J. Bacteriol.* **2019**, *201*. [CrossRef] [PubMed]
38. Todd, A.K.; Johnston, M.; Neidle, S. Highly prevalent putative quadruplex sequence motifs in human DNA. *Nucleic Acids Res.* **2005**, *33*, 2901–2907. [CrossRef]
39. Huppert, J.L.; Balasubramanian, S. Prevalence of quadruplexes in the human genome. *Nucleic Acids Res.* **2005**, *33*, 2908–2916. [CrossRef]
40. Eddy, J.; Maizels, N. Gene function correlates with potential for G4 DNA formation in the human genome. *Nucleic Acids Res.* **2006**, *34*, 3887–3896. [CrossRef] [PubMed]
41. Bedrat, A.; Lacroix, L.; Mergny, J.L. Re-evaluation of G-quadruplex propensity with G4Hunter. *Nucleic Acids Res.* **2016**. [CrossRef]
42. Brázda, V.; Kolomazník, J.; Lýsek, J.; Bartas, M.; Fojta, M.; Šťastný, J.; Mergny, J.-L. G4Hunter web application: A web server for G-quadruplex prediction. *Bioinformatics* **2019**, *35*, 3493–3495. [CrossRef] [PubMed]
43. Finan, T.M. The divided bacterial genome: Structure, function, and evolution. *Microbiol. Mol. Biol. Rev.* **2017**, *81*, e00019-17.
44. Yadav, V.K.; Abraham, J.K.; Mani, P.; Kulshrestha, R.; Chowdhury, S. QuadBase: Genome-wide database of G4 DNA-occurrence and conservation in human, chimpanzee, mouse and rat promoters and 146 microbes. *Nucleic Acids Res.* **2008**, *36*, D381–D385. [CrossRef] [PubMed]
45. Waller, Z.A.; Pinchbeck, B.J.; Buguth, B.S.; Meadows, T.G.; Richardson, D.J.; Gates, A.J. Control of bacterial nitrate assimilation by stabilization of G-quadruplex DNA. *Chem. Commun.* **2016**, *52*, 13511–13514. [CrossRef] [PubMed]
46. Rawal, P.; Kummarasetti, V.B.R.; Ravindran, J.; Kumar, N.; Halder, K.; Sharma, R.; Mukerji, M.; Das, S.K.; Chowdhury, S. Genome-wide prediction of G4 DNA as regulatory motifs: Role in *Escherichia coli* global regulation. *Genome Res.* **2006**, *16*, 644–655. [CrossRef]
47. Brázda, V.; Lýsek, J.; Bartas, M.; Fojta, M. Complex Analyses of Short Inverted Repeats in All Sequenced Chloroplast DNAs. *BioMed Res. Int.* **2018**, *2018*, 1097018. [CrossRef] [PubMed]
48. Čechová, J.; Lýsek, J.; Bartas, M.; Brázda, V. Complex analyses of inverted repeats in mitochondrial genomes revealed their importance and variability. *Bioinformatics* **2018**, *34*, 1081–1085. [CrossRef] [PubMed]
49. Cahoon, L.A.; Seifert, H.S. Transcription of a cis-acting, noncoding, small RNA is required for pilin antigenic variation in *Neisseria gonorrhoeae*. *PLoS Pathog.* **2013**, *9*, e1003074. [CrossRef] [PubMed]
50. Neidle, S. The structures of quadruplex nucleic acids and their drug complexes. *Curr. Opin. Struct. Biol.* **2009**, *19*, 239–250. [CrossRef]
51. Dhapola, P.; Chowdhury, S. QuadBase2: Web server for multiplexed guanine quadruplex mining and visualization. *Nucleic Acids Res.* **2016**, *44*, W277–W283. [CrossRef]
52. Sayers, E.W.; Agarwala, R.; Bolton, E.E.; Brister, J.R.; Canese, K.; Clark, K.; Connor, R.; Fiorini, N.; Funk, K.; Hefferon, T.; et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **2019**, *47*, D23–D28. [CrossRef] [PubMed]
53. Brázda, V.; Kolomazník, J.; Lýsek, J.; Hároníková, L.; Coufal, J.; Šťastný, J. Palindrome analyser—A new web-based server for predicting and evaluating inverted repeats in nucleotide sequences. *Biochem. Biophys. Res. Commun.* **2016**, *478*, 1739–1745. [CrossRef] [PubMed]
54. Computational Tools—Pandas 0.25.1 Documentation. Available online: https://pandas.pydata.org/pandas-docs/stable/user_guide/computation.html (accessed on 16 October 2019).
55. Suzuki, R.; Shimodaira, H. Pvclust: An R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* **2006**, *22*, 1540–1542. [CrossRef]
56. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403–410. [CrossRef]
57. Grant, C.E.; Bailey, T.L.; Noble, W.S. FIMO: Scanning for occurrences of a given motif. *Bioinformatics* **2011**, *27*, 1017–1018. [CrossRef] [PubMed]
58. Bailey, T.L.; Boden, M.; Buske, F.A.; Frith, M.; Grant, C.E.; Clementi, L.; Ren, J.; Li, W.W.; Noble, W.S. MEME SUITE: Tools for motif discovery and searching. *Nucleic Acids Res.* **2009**, *37*, W202–W208. [CrossRef]
59. Gertz, E.M.; Yu, Y.-K.; Agarwala, R.; Schäffer, A.A.; Altschul, S.F. Composition-based statistics and translated nucleotide searches: Improving the TBLASTN module of BLAST. *BMC Biol.* **2006**, *4*, 41. [CrossRef]

60. Wernersson, R. Virtual Ribosome—A comprehensive DNA translation tool with support for integration of sequence feature annotation. *Nucleic Acids Res.* **2006**, *34*, W385–W388. [[CrossRef](#)]
61. Artimo, P.; Jonnalagedda, M.; Arnold, K.; Baratin, D.; Csardi, G.; De Castro, E.; Duvaud, S.; Flegel, V.; Fortier, A.; Gasteiger, E. ExpPASy: SIB bioinformatics resource portal. *Nucleic Acids Res.* **2012**, *40*, W597–W603. [[CrossRef](#)]
62. Marchler-Bauer, A.; Derbyshire, M.K.; Gonzales, N.R.; Lu, S.; Chitsaz, F.; Geer, L.Y.; Geer, R.C.; He, J.; Gwadz, M.; Hurwitz, D.I.; et al. CDD: NCBI’s conserved domain database. *Nucleic Acids Res.* **2015**, *43*, D222–D226. [[CrossRef](#)]
63. Čutová, M.; Manta, J.; Porubiaková, O.; Kaura, P.; Šťastný, J.; Jagelská, E.B.; Goswami, P.; Bartas, M.; Brázda, V. Divergent distributions of inverted repeats and G-quadruplex forming sequences in *Saccharomyces cerevisiae*. *Genomics* **2020**, *112*, 1897–1901. [[CrossRef](#)] [[PubMed](#)]
64. Guo, J.U.; Bartel, D.P. RNA G-quadruplexes are globally unfolded in eukaryotic cells and depleted in bacteria. *Science* **2016**, *353*. [[CrossRef](#)] [[PubMed](#)]
65. Galtier, N.; Tourasse, N.; Gouy, M. A nonhyperthermophilic common ancestor to extant life forms. *Science* **1999**, *283*, 220–221. [[CrossRef](#)]
66. Klein, R.J.; Misulovin, Z.; Eddy, S.R. Noncoding RNA genes identified in AT-rich hyperthermophiles. *Proc. Natl. Sci. Acad. USA* **2002**, *99*, 7542–7547. [[CrossRef](#)] [[PubMed](#)]
67. Lyons, S.M.; Gudanis, D.; Coyne, S.M.; Gdaniec, Z.; Ivanov, P. Identification of functional tetramolecular RNA G-quadruplexes derived from transfer RNAs. *Nat. Commun.* **2017**, *8*, 1127. [[CrossRef](#)]
68. Gebetsberger, J.; Zywicki, M.; Künzi, A.; Polacek, N. tRNA-derived fragments target the ribosome and function as regulatory non-coding RNA in *Haloflexax volcanii*. *Archaea* **2012**, *2012*, 260909. [[CrossRef](#)]
69. Magnus, M.; Kappel, K.; Das, R.; Bujnicki, J.M. RNA 3D structure prediction guided by independent folding of homologous sequences. *BMC Bioinf.* **2019**, *20*, 512. [[CrossRef](#)]
70. Kamura, T.; Katsuda, Y.; Kitamura, Y.; Ihara, T. G-quadruplexes in mRNA: A key structure for biological function. *Biochem. Biophys. Res. Commun.* **2020**. [[CrossRef](#)]
71. Qu, Z.; Adelson, D.L. Evolutionary conservation and functional roles of ncRNA. *Front. Genet.* **2012**, *3*. [[CrossRef](#)]
72. Buddeweg, A.; Daume, M.; Randau, L.; Schmitz, R.A. Noncoding RNAs in Archaea: Genome-Wide Identification and Functional Classification. *Meth. Enzymol.* **2018**, *612*, 413–442. [[CrossRef](#)]
73. Luo, H.; Gao, F. DoriC 10.0: An updated database of replication origins in prokaryotic genomes including chromosomes and plasmids. *Nucleic Acids Res.* **2019**, *47*, D74–D77. [[CrossRef](#)] [[PubMed](#)]
74. Cossu, M.; Da Cunha, V.; Toffano-Nioche, C.; Forterre, P.; Oberto, J. Comparative genomics reveals conserved positioning of essential genomic clusters in highly rearranged Thermococcales chromosomes. *Biochimie* **2015**, *118*, 313–321. [[CrossRef](#)] [[PubMed](#)]
75. Matsunaga, F.; Forterre, P.; Ishino, Y.; Myllykallio, H. In vivo interactions of archaeal Cdc6/Orc1 and minichromosome maintenance proteins with the replication origin. *Proc. Natl. Acad. Sci. USA* **2001**, *98*, 11152–11157. [[CrossRef](#)] [[PubMed](#)]
76. Dueber, E.C.; Costa, A.; Corn, J.E.; Bell, S.D.; Berger, J.M. Molecular determinants of origin discrimination by Orc1 initiators in archaea. *Nucleic Acids Res.* **2011**, *39*, 3621–3631. [[CrossRef](#)]
77. Norais, C.; Hawkins, M.; Hartman, A.L.; Eisen, J.A.; Myllykallio, H.; Allers, T. Genetic and physical mapping of DNA replication origins in *Haloflexax volcanii*. *PLoS Genet.* **2007**, *3*, e77. [[CrossRef](#)]
78. Wu, Z.; Liu, J.; Yang, H.; Liu, H.; Xiang, H. Multiple replication origins with diverse control mechanisms in *Haloarcula hispanica*. *Nucleic Acids Res.* **2013**, *42*, 2282–2294. [[CrossRef](#)]
79. Zhuang, X.; Tang, J.; Hao, Y.; Tan, Z. Fast detection of quadruplex structure in DNA by the intrinsic fluorescence of a single-stranded DNA binding protein. *J. Mol. Recognit.* **2007**, *20*, 386–391. [[CrossRef](#)]
80. Mendoza, O.; Bourdoncle, A.; Boulé, J.-B.; Brosh, R.M.; Mergny, J.-L. G-quadruplexes and helicases. *Nucleic Acids Res.* **2016**, *44*, 1989–2006. [[CrossRef](#)]
81. Beaume, N.; Pathak, R.; Yadav, V.K.; Kota, S.; Misra, H.S.; Gautam, H.K.; Chowdhury, S. Genome-wide study predicts promoter-G4 DNA motifs regulate selective functions in bacteria: Radioresistance of *D. radiodurans* involves G4 DNA-mediated regulation. *Nucleic Acids Res.* **2013**, *41*, 76–89. [[CrossRef](#)]
82. Gehring, K.; Leroy, J.-L.; Guéron, M. A tetrameric DNA structure with protonated cytosine-cytosine base pairs. *Nature* **1993**, *363*, 561–565. [[CrossRef](#)]

83. Barrett, T.; Clark, K.; Gevorgyan, R.; Gorelenkov, V.; Gribov, E.; Karsch-Mizrachi, I.; Kimelman, M.; Pruitt, K.D.; Resenchuk, S.; Tatusova, T.; et al. BioProject and BioSample databases at NCBI: Facilitating capture and organization of metadata. *Nucleic Acids Res.* **2012**, *40*, D57–D63. [[CrossRef](#)] [[PubMed](#)]
84. Bartas, M.; Brázda, V.; Karlický, V.; Červeň, J.; Pečinka, P. Bioinformatics analyses and in vitro evidence for five and six stacked G-quadruplex forming sequences. *Biochimie* **2018**, *150*, 70–75. [[CrossRef](#)]
85. Risitano, A.; Fox, K.R. Stability of Intramolecular DNA Quadruplexes: Comparison with DNA Duplexes. *Biochemistry* **2003**, *42*, 6507–6513. [[CrossRef](#)]
86. Couturier, M.; Gabelle, D.; Forterre, P.; Nadal, M.; Garnier, F. The reverse gyrase TopR1 is responsible for the homeostatic control of DNA supercoiling in the hyperthermophilic archaeon *Sulfolobus solfataricus*. *Mol. Microbiol.* **2020**, *113*, 356–368. [[CrossRef](#)] [[PubMed](#)]
87. Chambers, V.S.; Marsico, G.; Boutell, J.M.; Di Antonio, M.; Smith, G.P.; Balasubramanian, S. High-throughput sequencing of DNA G-quadruplex structures in the human genome. *Nat. Biotechnol.* **2015**, *33*, 877. [[CrossRef](#)]
88. Hänsel-Hertsch, R.; Spiegel, J.; Marsico, G.; Tannahill, D.; Balasubramanian, S. Genome-wide mapping of endogenous G-quadruplex DNA structures by chromatin immunoprecipitation and high-throughput sequencing. *Nat. Protoc.* **2018**, *13*, 551. [[CrossRef](#)] [[PubMed](#)]
89. Hänsel-Hertsch, R.; Di Antonio, M.; Balasubramanian, S. DNA G-quadruplexes in the human genome: Detection, functions and therapeutic potential. *Nat. Rev. Mol. Cell. Biol.* **2017**, *18*, 279. [[CrossRef](#)]



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

G-quadruplexes in helminth parasites

Alessio Cantara^{1,2,†}, Yu Luo^{3,4,†}, Michaela Dobrovolná^{5,†}, Natalia Bohalova^{1,2},
Miroslav Fojta¹, Daniela Verga^{3,6}, Lionel Guittat^{4,7}, Anne Cucchiari⁴,
Solène Savrimoutou⁸, Cécile Häberli^{9,10}, Jean Guillon⁸, Jennifer Keiser^{9,10},
Václav Brázda^{1,5,*} and Jean Louis Mergny^{1,4,*}

¹Institute of Biophysics, Czech Academy of Sciences, Královopolská 135, 612 65 Brno, Czech Republic,

²Department of Experimental Biology, Faculty of Science, Masaryk University, Kamenice 5, 62500 Brno, Czech Republic, ³CNRS UMR9187, INSERM U1196, Université Paris-Saclay, F-91405 Orsay, France, ⁴Laboratoire d'Optique et Biosciences, Ecole Polytechnique, CNRS, INSERM, Institut Polytechnique de Paris, 91128 Palaiseau, France, ⁵Faculty of Chemistry, Brno University of Technology, Purkyňova 118, 612 00 Brno, Czech Republic, ⁶CNRS UMR9187, INSERM U1196, Institut Curie, PSL Research University, F-91405 Orsay, France, ⁷Université Sorbonne Paris Nord, UFR SMBH, Bobigny, France, ⁸ARNA Laboratory, Université de Bordeaux, INSERM U1212, CNRS UMR 5320, UFR des Sciences Pharmaceutiques, Bordeaux, France, ⁹Department of Medical Parasitology and Infection Biology, Swiss Tropical and Public Health Institute, Basel, Switzerland and ¹⁰University of Basel, Basel, Switzerland

Received June 30, 2021; Revised February 07, 2022; Editorial Decision February 08, 2022; Accepted February 25, 2022

ABSTRACT

Parasitic helminths infecting humans are highly prevalent infecting ~2 billion people worldwide, causing inflammatory responses, malnutrition and anemia that are the primary cause of morbidity. In addition, helminth infections of cattle have a significant economic impact on livestock production, milk yield and fertility. The etiological agents of helminth infections are mainly Nematodes (roundworms) and Platyhelminths (flatworms). G-quadruplexes (G4) are unusual nucleic acid structures formed by G-rich sequences that can be recognized by specific G4 ligands. Here we used the G4Hunter Web Tool to identify and compare potential G4 sequences (PQS) in the nuclear and mitochondrial genomes of various helminths to identify G4 ligand targets. PQS are nonrandomly distributed in these genomes and often located in the proximity of genes. Unexpectedly, a Nematode, *Ascaris lumbricoides*, was found to be highly enriched in stable PQS. This species can tolerate high-stability G4 structures, which are not counter selected at all, in stark contrast to most other species. We experimentally confirmed G4 formation for sequences found in four different parasitic helminths. Small molecules able to selectively recognize G4 were found to bind to *Schistosoma mansoni* G4 motifs. Two of these ligands demonstrated po-

tent activity both against larval and adult stages of this parasite.

INTRODUCTION

Helminth infections caused by parasitic Nematodes (roundworms) and Platyhelminths (flatworms) are among the most prevalent afflictions for people living in poor areas of the world with over a quarter of the total human population affected worldwide (1,2). Parasitic worm infections cause anemia, malnutrition, allergies, bloody diarrhea, bowel cramps and inflammation associated with colonic polyposis (3) and increased susceptibility to HIV and progression to AIDS, resulting in many obstructive pathologies (3,4). In addition, severe anemia in pregnancy is associated with neonatal prematurity and reduced birthweight (3,5). Cattle infections have a significant economic impact on livestock production due to a reduction in growth, milk yield and fertility (6). Recent estimates suggest that *Ascaris lumbricoides* infects over a billion, *Trichuris trichiura* 795 million (7), and *Strongyloides stercoralis* 30–100 million people (8). Hookworms such as *Necator americanus* and *Ancylostoma duodenale* cause hookworm diseases, which are associated with blood loss and anemia. Schistosomiasis is the third most reported global tropical disease caused by trematode flukes of the genus *Schistosoma* (9). The three most common species of parasitic trematodes of the family Schistosomatidae are *Schistosoma mansoni*, *Schistosoma japonicum* and *Schistosoma haematobium* (1,9). The greatest numbers of *S. mansoni* infections occur in Sub-Saharan

*To whom correspondence should be addressed. Tel: +33 766290967; Email: jean-louis.mergny@inserm.fr
Correspondence may also be addressed to Václav Brázda. Email: vabdna@gmail.com

†The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.

Africa, the Middle East, the Caribbean, Brazil, Venezuela and Suriname. *S. japonicum* is localized to Asia, primarily the Philippines, Indonesia, and China. *S. haematobium* is highly prevalent in Sub-Saharan Africa, the Middle East, and was recently reported in Corsica (France) (2). Schistosomiasis is estimated to affect ~200 million people worldwide (10). Few treatments are available to fight worm infections. In addition, widespread use of drugs such as benzimidazoles selects for drug-resistant parasite strains (11), and reduced praziquantel efficacy leads to low egg reduction rates (12). The recent sequencing of the genome of helminths (e.g. *S. mansoni*, *S. japonicum* and *S. haematobium*) (13–15) offers new opportunities to identify novel targets in key genes of the parasites.

Identifying potential G4-forming sequences (PQS) in key genes associated with the infectivity of parasitic worms should allow the prediction of new druggable targets for G4 ligands. G4 are noncanonical nucleic acid secondary structures (16) formed by G-rich sequences that are built of stacked tetrads (also called ‘G-quartets’) constituted of four Hoogsteen hydrogen-bonded guanine bases. They can display a wide variety of topologies, resulting from various possible combinations of DNA/RNA strand directions, as well as variations in loop size and molecularity (17). G4 are further stabilized by the presence of monovalent cations such as potassium or sodium, which are abundant in cells (18). G4 can be found throughout a variety of genomes (19), and are overrepresented in the promoter regions of genes, especially regulatory genes involved in cell proliferation or survival (20,21) as well as in regions which regulate important biological processes (22) including immune response, transcriptional activation, DNA damage repair (23) and telomere maintenance (24), as shown mostly in mammalian cells. The interest in the therapeutic potential of gene promoters containing G4 has resulted in a rapidly increasing number of studies during the past decade in which small molecules have been used to act as G4 stabilizers, with reports of transcription inhibition in cell-based assays.

G4 structures have only been investigated in a few Platyhelminths and Nematode species. For instance, 1,500 PQS were found in *Caenorhabditis elegans* using the *Quadparser* algorithm, of which ~500 are associated with transcription start site regions. The peak in PQS density coincides almost exactly with the nucleosome depleted region, which is consistent with the hypothesis that functional PQS may be located outside nucleosome bound regions (25). G4 stable under physiological conditions (K⁺-rich buffer) and stabilized by Pyridostatin (PDS, a specific G4-targeting small molecule) were identified in *C. elegans* (26). The presence of G4 structures in heterochromatin and the difference in G4 staining between somatic and stem cells in germline DNA of the flatworm *Macrostomum lignano* pointed out the possibility that the resolution or suppression of G4 structures is important for stem cells with regenerative potential (27). A study conducted on *A. lumbricoideis* shows that the fluorescent G4 ligand Q1 binds selectively to the antiparallel telomeric G4 (28). Very recently, G4 motifs were identified in the genome of *S. mansoni* and the authors confirmed the presence of G4 in adult worms, by means of the BG4 G-quadruplex-specific antibody (29).

Given the lack of information available in regard to the nature of G4s formed in infective species of Platyhelminths and Nematodes, we analyzed differences in potential G4-forming sequences (PQS) presence, frequency and localization in the genomes of six Platyhelminths and four Nematodes using the G4Hunter tool. Comparing the genomes of parasites from these two phyla with different infectivity (parasitic and non-parasitic worms) may reveal details of processes driving pathogenicity and new possible drug targets. Finally, we showed that the recently-identified G4 motifs found in *S. mansoni* (29) may be targeted by G4 ligands, leading to anti-parasitic effects.

MATERIALS AND METHODS

Source of DNA sequences

The set of selected genomes (both mitochondrial and nuclear DNAs) of two Clades (Platyhelminths and Nematodes) belonging to the Clade Nephzoa (subgroup Protostomia) was downloaded from the Genome database of the NCBI. Six Platyhelminths (*Schistosoma haematobium*, *Schistosoma japonicum*, *Schistosoma mansoni*, *Trichobilharzia regenti*, *Dibothriocephalus latus* and *Taenia asiatica*), and four Nematoda (*Strongyloides stercoralis*, *Trichuris trichiura*, *Ascaris lumbricoideis* and *Caenorhabditis elegans*) were chosen for an initial analysis. Four additional nematode species (*Anisakis simplex*, *Ascaris suum*, *Parascaris equorum* and *Toxocara canis*) were added for comparison with *Ascaris lumbricoideis*. All accession numbers are provided in Supplementary Material 01.

Analytical process

All sequences were analyzed using the G4Hunter Web tool (30) (<http://bioinformatics.ibp.cz/#/analyse/quadruplex>) which can read National Center for Biotechnology Information (NCBI) IDs. Unless specified otherwise, parameters for G4Hunter were set to 25 nucleotides for window size and ≥ 1.2 for threshold scores. This threshold appears as a reasonable compromise, giving few false positives (sequences not forming a G4 despite a G4Hunter score above threshold) and false negatives (sequences able to form a stable G4 despite having a G4Hunter score below threshold). Scores above 1.2 correspond to sequences having a higher guanine content and likely to form stable G4s. To rank sequences based on score, motifs were binned in five intervals covering the G4Hunter scores 1.2–1.4, 1.4–1.6, 1.6–1.8, 1.8–2.0 and >2.0 . Data were merged in a single Excel file where was made the statistical evaluation and are accessible in Supplementary Material 02.

To test whether the PQS occurrence in chromosomal breakpoint regions is significantly different than in a randomly shuffled sequence, we generated 40 random sequences with length and nucleotide content the same as in original *A. suum* breakpoints. All sequences were randomly shuffled using the Sequence Manipulation Suite (https://www.bioinformatics.org/sms2/shuffle_dna.html) and were manually analyzed using the G4Hunter Web tool (30). The parameters for G4Hunter were set to 25 nucleotides for window size and thresholds of 1.2 or more (several thresholds

were considered). Sequences were merged in a single Excel file where was made the statistical evaluation using *t*-test.

LOGO sequences

NCBI sequences in FASTA format were downloaded and the dataset was uploaded to SnapGene program. For every PQS we used the corresponding sequences from all analyzed genomes and alignments were generated using the Clustal Omega tool. All PQS found were searched in aligned sequences and WebLogo 3 was used for generating LOGO sequences Supplementary Material 04.

Statistical evaluation

Raw data were converted in .xlsx file format and analyzed through Microsoft Excel. All data files are available in Supplementary Materials. Correlation was evaluated by the Spearman's rank correlation coefficient (r_s) and are presented in Supplementary Material 05.

Samples

Oligonucleotides were purchased from Eurogentec (Belgium) and used without further purification. Stock solutions were prepared at 100 μ M strand concentration for the unlabeled oligonucleotides and at 200 μ M strand concentration for double-labeled oligonucleotides in ddH₂O. Sequences of tested G4 motifs and control G4s, single-strands, and duplexes are shown in Supplementary Table S5. All oligonucleotides were annealed (95°C for 5 min and slowly cooled to room temperature) in the corresponding buffer before measurements.

UV-melting assay

3 μ M oligonucleotide solutions were annealed in K100 buffer (100 mM KCl, 10 mM lithium cacodylate, 90 mM LiCl, pH 7.2). UV-melting profiles were recorded with a Cary 300 spectrophotometer (Agilent Technologies, France). Heating runs were performed between 10°C and 95°C, the temperature was increased by 0.2°C/min, and absorbance was recorded at 260 and 295 nm (31).

Circular dichroism

3 μ M oligonucleotide solutions were annealed in K100 buffer. CD spectra were recorded on a J-1500 spectropolarimeter (JASCO, France) at room temperature (25°C), using a scan range of 300–220 nm, a scan rate of 100 nm/min and averaging four accumulations.

FRET-melting assay

FRET melting assay was performed in 96-well plates and the fluorescence of dual-labeled G4-forming oligonucleotides (including F21T; sequences are shown in Supplementary Table S7) was recorded using a CFX96 qPCR instrument (Biorad). F21T sequence was annealed at 0.23 μ M in K10 buffer (10 mM KCl, 10 mM lithium cacodylate, 90

mM LiCl, pH 7.2), then the oligonucleotide was added to each well (final strand concentration of 0.2 μ M) which was incubated with or without the tested ligands at 2 and 5 μ M final concentration, to a final volume of 25 μ l. Competition experiments were performed in the presence of non-labeled sequences, including one auto complementary duplex, ds26, one parallel G4, c-myc and G4s from *S. Mansoni*. The microplate was incubated at 25°C for 5 min, after which the temperature was increased by increments of 0.5°C/min to reach 95°C. The collected signal was normalized to 1 and the melting temperature (T_m) was defined when the normalized signal was 0.5. ΔT_m corresponds to the difference of T_m between the oligonucleotides with and without the ligands. This FRET-melting assay was done in duplicates.

FRET-melting competition assay

FRET melting competition (FRET-MC) experiments were performed in 96-well plates using a HT7900 RT-PCR instrument (Applied BioSystem), as previously described (32). 50 μ M oligonucleotide solutions were annealed in K10 buffer. Each well contained 3 μ M competitors, 0.2 μ M fluorescent oligonucleotide F21T in the presence or absence of 0.4 μ M G4 ligand (PhenDC3) in K10 buffer, for a final volume of 25 μ l. Samples were kept at 25°C for 5 min, then the temperature was increased by 0.5°C per minute until 95°C, and the FAM channel was used to collect the fluorescence signal. The T_m of an oligonucleotide is defined as the temperature at which 50% of the oligonucleotide is unfolded. ΔT_m is determined as the difference in T_m with the sample containing F21T in the absence of PhenDC3. Each experimental condition was tested in duplicate on two separate plates.

Thermal difference spectra (TDS) and Isothermal difference spectra (IDS)

Absorbance spectra were recorded on a Cary 300 spectrophotometer (Agilent Technologies, France) (scan range: 500–200 nm; scan rate: 600 nm/min; automatic baseline correction).

- **TDS:** 3 μ M oligonucleotide solutions were annealed in K100 buffer. After recording the first spectra (folded) at 25°C, temperature was increased to 95°C, and the second UV-absorbance spectra was recorded after 15 min of equilibration at high temperature. TDS corresponds to the arithmetic difference between the initial (folded; 25°C) and second (unfolded; 95°C) spectra (33).
- **IDS:** 3 μ M oligonucleotide solutions were annealed in Li-Caco10 buffer (10 mM lithium cacodylate, pH 7.2). Absorbance spectra were first recorded at 25°C in the absence of any stabilizing cation. 1 M KCl was added after recording the first spectrum, to a final potassium concentration of 100 mM KCl. The second UV-absorbance spectrum was recorded after 15 min of equilibration. IDS correspond to the arithmetic difference between the initial (unfolded) and final (folded, thanks to the addition of K⁺) spectra, after correction for dilution.

G-Quadruplex fluorescent light-up probes

- **ThT** (Thioflavin T) was used as previously described (34). 7.5 μ M oligonucleotide solutions were annealed in K100 buffer. Each component was added in the order: 10 μ l K100 buffer, 10 μ l oligonucleotide and 5 μ l of 10 μ M ThT (dissolved in milli-Q water). The plate was shaken for 5 min and was incubated for 10 min at room temperature. Fluorescence intensity was collected at 490 nm after excitation at 420 nm in a TECAN M1000 pro plate reader.
- **NMM** (*N*-methyl mesoporphyrin IX) was used under the same condition as ThT, except that fluorescence intensity was collected at 610 nm after excitation at 380 nm in a TECAN M1000 pro plate reader.

G-Quadruplex ligands

The synthesis of G4 ligands tested against *S. mansoni* was previously described (35–37). Stock solutions were prepared at 10 mM in DMSO.

Antischistosomal activity of G-quadruplex ligands

Newly transformed schistosomula (NTS) drug assay. *S. mansoni* cercariae were collected from infected snails and mechanically transformed to newly transformed schistosomula (NTS). 30–40 NTS/well were incubated with the drugs for 72 h at 37°C, 5% CO₂ in a final well volume of 200–250 μ l. Compounds were tested in triplicate and the highest concentration of DMSO (<1%) served as control. Evaluation was done by microscopic readout (Carl Zeiss, Germany, magnification 80x) as summarized in a previous publication (38).

Adult S. mansoni drug assay. Animal studies were carried out following Swiss national and cantonal regulations on animal welfare at the Swiss Tropical and Public Health Institute (Basel, Switzerland (Swiss TPH) (permission no. 2070). Female mice (NMRI; age 3 weeks; weight *ca.* 20–22 g) were purchased from Charles River, Germany. Mice were kept under environmentally controlled conditions (temperature ~25°C; humidity ~70%; 12 h light and 12 h dark cycle) with free access to water and rodent diet, and acclimatized for 1 week before infection.

Adult schistosomes were collected by mechanical picking from the hepatic portal system and mesenteric veins of mice 49-day post-infection with 100 *S. mansoni* cercariae. Worms were incubated with the compounds for 72 h. Wells with 1% DMSO served as negative controls. IC₅₀ values were calculated using CalcuSyn Version 2.0 (Biosoft, Cambridge, UK). Phenotypes were evaluated under an inverted microscope and viability scores calculated (38).

RESULTS

We have selected 10 helminth organisms based on their impact on health or relevance as model species. From the accessible genomes, three of the most pathogenic species of Platyhelminths (*Schistosoma haematobium*, *Schistosoma japonicum* and *Schistosoma mansoni*) and three important Nematode species (*Strongyloides stercoralis*, *Trichuris*

trichiura, and *Ascaris lumbricoides*) were selected. As reference organisms we have additionally selected three Platyhelminths (*Trichobilharzia regenti*, *Dibothriocephalus latus*, *Taenia asiatica*) and *Caenorhabditis elegans*, one of the best studied Nematodes. Both mitochondrial DNA (mtDNA) and nuclear genomes were analyzed by the G4Hunter algorithm for the presence of PQS.

PQS in mitochondrial DNA

At first, we analyzed the mitochondrial genomes of the 10 helminths listed above (Supplementary Table S1). mtDNA length varied from 13,608 to 15,003 bp, with a GC content between 23% (*S. stercoralis*) and 32% (*T. trichiura*). The results show that GC content is poorly correlated with the number of detected PQS: we found only one PQS sequence for the organism with the highest GC content (*T. trichiura*) while *S. stercoralis*, which has the mitochondrial genome with the lowest GC content in our dataset, has 6 PQS in its mtDNA. In total, we found 77 PQS with a G4Hunter score above 1.2, but none with a score above 1.4 (in other words, all motifs found are in the 1.2–1.4 interval). A 1.2 threshold is considered as a reasonable compromise to identify G4 prone motif (45); higher scores correspond to motifs capable of forming very stable quadruplexes. The majority of these sequences (40/77) were found in the mtDNA of a single species, *T. regenti*.

To analyze the PQS localization in mtDNA, we downloaded their annotations from NCBI and overlaid the PQS presence with these features (Supplementary Material 03). The organism with the highest number of PQS in mtDNA, *T. regenti*, has the majority of PQS in the repeat region of its mtDNA. However, PQS were also found in gene regions of this organism encoding cytochrome *c* oxidase subunit III, cytochrome *b*, NADH dehydrogenases subunits 4 and 2 and ATP synthase F0 subunit 6.

For *S. mansoni*, all (11 out of 11) PQS are located in the CDS of the genes coding for cytochrome *c* oxidase and NADH dehydrogenase subunit. This can be found in Supplementary Material 03 and Supplementary Table S1. The results demonstrate a nonrandom distribution of PQS in mtDNA of analyzed organisms according to PQS position and reveal that the prevalence of PQS is related mainly to production of NADH dehydrogenase subunits in *S. mansoni*. Supplementary Table S2 shows that *S. mansoni* is the only organism with an overlap of the PQS with the gene coding for the NADH dehydrogenase subunits 2, 5 and 4; the other two PQS are located inside the region coding for the cytochrome *c* oxidase subunits I and II.

To explore if the PQS are over-represented in regions coding NADH dehydrogenase subunits, we took feature tables containing annotations of known features found in mtDNA sequences and counted them. We then counted features with PQS. As mentioned before, the majority of the PQS is located inside the mitochondrial genes and coding sequences (CDS). PQS within the region coding for NADH dehydrogenase subunits are present in over 17% of all NADH dehydrogenase subunit coding regions and over 36% of all regions coding for cytochrome *c* oxidase subunits. In contrast, only 10% of rRNA and <2% of tRNA sequences contained PQS suggesting enrichment in NADH dehydrogenase sub-

unit and cytochrome *c* oxidase subunits. Data are available in Supplementary Material 03. PQS within the region coding for cytochrome *c* oxidase subunits are present in each analyzed organism. The position of this PQS may vary; it is found either in the region coding for the subunit I (*S. haematobium*, *S. japonicum* and *S. mansoni*) or subunit II (*S. mansoni*). A complete table can be found in Supplementary Material 03.

Predicted PQS in mtDNA of the three analyzed schistosoma species were compared looking for conserved motifs or sequences. We aligned the predicted PQS and generated their LOGO using the WEB LOGO tool (39) and the results are presented in Supplementary Material 04 for the most conserved sequences among the three *Schistosoma* species analyzed here: *S. haematobium*, *S. japonicum* and *S. mansoni*. A conservation of ~70% is found in these motifs. A similar level of conservation was seen among all six Platyhelminths. Evaluation of the PQS position shows that these most conserved sequences are in the COX1 gene.

PQS in nuclear DNA

Using standard values for G4Hunter (*i.e.* a window size of 25 nucleotides and threshold score of 1.2), we found over 1.3 million PQS among all 10 genomes. Overall, we do not advocate the use of a single threshold, but prefer to analyze data for various G4Hunter score windows. For this reason, we present these with different threshold windows (Table 1). We do have a good idea on false positive (FP) rate depending on threshold (<5% for scores >1.25; <2% for score >1.5 and close to 0 for scores >1.75; (45) and unpublished data; false positives are defined as sequences predicted to form a stable quadruplex, but cannot be confirmed experimentally). Our understanding on false negative (FN) rate as a function of threshold is less reliable, primarily because we did not explore as many sequences with relatively low G4Hunter scores (false negatives are sequences forming stable G4 which are missed by the algorithm). Our unpublished data with Dr Laurent Lacroix suggests that 1.2 is a reasonable threshold to maximize accuracy (*i.e.* minimize the fraction of FP + FN). For this reason, we chose this value as the minimal threshold considered. Higher thresholds (up to 2.0) tend to select sequences with high stability and propensity to form G4 structures.

Total PQS counts, percentage of GC and PQS frequencies characteristics for each organism are summarized in Table 2. The length of analyzed nuclear genomes varied from 42 Mbp (*S. stercoralis*) to 701 Mbp (*T. regenti*). The mean GC content was 34.9%, with a minimum of 22.1% for *S. stercoralis*, and a maximum of 42.2% for *T. trichiura*. The highest PQS frequencies were found in *D. latus* in which 473 thousand PQS were present in a 531 Mb genome (giving a PQS frequency of 0.89 PQS per 1000 nucleotides) and *T. asiatica* with 151,000 PQS in a 168 Mb genome, giving a similar PQS frequency of 0.90 PQS per kb (exact values provided in Table 2). In contrast, the parasite with the lowest PQS frequency was *S. stercoralis* with only 3,037 PQS for a 42.7 Mbp nuclear genome, giving a PQS frequency of 0.071 PQS per kb. We then compared the PQS frequency between moderately and highly infective parasites (40,41). Interestingly,

Table 1. Total number of PQS and their frequencies according to their G4Hunter score

Interval of G4Hunter score	Number of PQS in dataset	PQS frequency per 1000 bp
All		
1.2–1.4	1,281,682	0.396
1.4–1.6	29,381	0.010
1.6–1.8	5,631	0.002
1.8–2.0	2,158	0.001
2.0 or more	2,346	0.001
Platyhelminths		
1.2–1.4	1,051,587	0.440
1.4–1.6	16,715	0.007
1.6–1.8	976	≈0
1.8–2.0	42	≈0
2.0 or more	24	≈0
Nematodes		
1.2–1.4	230,095	0.329
1.4–1.6	12,666	0.014
1.6–1.8	4,655	0.004
1.8–2.0	2,116	0.002
2.0 or more	2,322	0.002

in these groups, the highest % of PQS was found in less infective Platyhelminths (with an average of 2.2 PQS per 1000 nt), followed by less infective Nematodes (1.6 PQS per kb (Table 2C). Detailed results are shown in Supplementary Material 02.

PQS frequency partially depends on GC content (a GC-poor genome is less likely to exhibit local G-rich motifs necessary for G4 formation). We present the global density in PQS (for all motifs with a G4Hunter score above 1.2) as a function of GC content for each individual organism in Figure 1A. A detailed analysis for each G4Hunter score interval is provided in Supplementary Material 06. The Spearman's rank correlation coefficient (r_s) was used to determine the association between PQS and GC content (Supplementary Material 05). PQS frequency is correlated with GC% ($r_s = 0.78$). This correlation is however far from perfect: for example, *T. trichiura* and *T. asiatica* have almost the same GC content, but their PQS frequencies differ considerably (0.263 for *T. trichiura* vs 0.881 PQS per kb for *T. asiatica*). When considering all potential G4s (for a threshold ≥ 1.2), Platyhelminths look enriched with PQS (with a total of 1,069,344 PQS found) compared to Nematoda (with a total of 251,854 PQS found).

The results appear completely different if we restrict the analysis of PQS frequency to motifs with a G4Hunter score above 1.6 (Figure 1B)—in this case most species exhibit very low PQS frequencies (<0.002/kb), with the striking exception of *A. lumbricoides*, for which the density is at least 10-fold higher than in any other species.

In contrast with mitochondria (where no PQS with a score >1.4 was found), we still found a significant number of PQS with high G4Hunter scores (Table 1). However, and as for other species including bacteria (42) and archaea (43), the number of G4 motifs found drops significantly when higher thresholds are selected. This observation is valid for all genomes tested so far, including viruses, bacteria, archaea and eukaryotes. The majority of the PQS have a score in the range 1.2–1.4: most PQS sequences have a relatively

Table 2. G4Hunter analysis results for nuclear DNA. (A) Statistics for all tested organisms and clades. Seq (number of sequences), Median, Short and Long correspond to the median, min. and max. lengths in the dataset, GC% (GC content), PQS (number of PQS), Mean *f* (mean of PQS per kb), Min *f* (lowest frequency), Max *f* (highest frequency). (B) Statistics for individual organisms. (C) Statistics for organisms divided into highly and moderately infective categories

(A)

Domain	Seq	Median	Short	Long	PQS	Mean <i>f</i>	Min <i>f</i>	Max <i>f</i>	GC%
All	10	346,434,783	42,674,647	701,762,036	1,321,198	0.409	0.071	0.897	34.9
Group	Seq	Median	Short	Long	PQS	Mean <i>f</i>	Min <i>f</i>	Max <i>f</i>	GC%
Platyhelminths	6	406,161,092	168,679,183	701,762,036	1,069,344	0.448	0.190	0.897	35.6
Nematodes	4	87,891,397	42,674,647	316,975,410	251,854	0.350	0.071	0.563	34.1

(B)

Organism	Length	PQS	Mean <i>f</i>	GC%
<i>Schistosoma haematobium</i>	375,894,156	71,575	0.190	32.11
<i>Schistosoma japonicum</i>	402,743,189	84,683	0.210	31.23
<i>Schistosoma mansoni</i>	409,579,008	88,178	0.215	34.66
<i>Trichobilharzia regenti</i>	701,762,036	200,260	0.285	33.34
<i>Dibothriocephalus latus</i>	531,434,409	473,393	0.891	40.07
<i>Taenia asiatica</i>	168,679,183	151,255	0.897	42.06
<i>Strongyloides stercoralis</i>	42,674,647	20,471	0.071	22.11
<i>Trichuris trichiura</i>	75,496,394	3,037	0.271	42.22
<i>Ascaris lumbricoides</i>	316,975,410	178,583	0.563	36.48
<i>Caenorhabditis elegans</i>	100,286,401	49,763	0.496	35.44

(C)

Infectiousness	Seq	PQS	Mean PQS	Mean <i>f</i>	Mean PQS%	Mean PQS/GC%	Mean GC%
Platyhelminths High (Schistosomas spp)	3	244,436	81,478.67	0.205	0.665	358,253.80	32.66
Platyhelminths Low	3	824,908	274,969.33	0.691	2.229	713,947.83	38.49
Nematoda Highly (all except <i>C. elegans</i>)	3	202,091	67,363.67	0.302	1.060	183,916.89	33.60

low G4Hunter score. PQS in the next interval (1.4–1.6) are less frequent, and the drop continues for higher values (note the logarithmic axis on the Y-scale; Figure 2).

Compared to mtDNA, the nuclear DNAs are still poorly annotated in these species, with exception of *C. elegans*. Contrary to mtDNA, where PQS are located in the repeat region and gene regions, PQS are slightly enriched before and after gene regions, in the annotations for various RNA such as ncRNA, tRNA and precursor_RNA. The most significant enrichment for PQS were found in the regions before and after rRNA and prim.transcript (Supplementary Material 10: *Elegans*_Annotations-result.xls).

The comparison between phyla is interesting; *Homo sapiens*, archaea and bacteria are provided for comparison (43). Of note, PQS with a high G4Hunter score are very rare in Platyhelminths, while low-score PQS are extremely abundant. Platyhelminths behave like archaea and bacteria, with a stronger counter selection against very stable G4 than in *Homo sapiens* (Figure 2). Very stable G4 are therefore strongly counter selected in Platyhelminths as compared to humans: these G4 with high G4Hunter scores are extremely rare.

We then analyzed individual Nematode species in Figure 2B. We plotted data for each species (*Homo sapiens* and the group of three nematodes are again provided for comparison). As can be shown, *A. lumbricoides* is the organism that maintains the highest relative level of stable PQS at high thresholds, while the three other Nematoda exhibit a drop in relative number of PQS that is comparable (for *C. elegans*) or even sharper (for *S. stercoralis* and *T. trichiura*) than *Homo sapiens*.

A. lumbricoides is therefore unique among nematodes, and actually unique among all species we have studied so far: while its overall density in all G4 motifs is not remarkable (Figure 2), it can tolerate high-stability G4 structures, suggesting that these motifs are not counter-selected at all. The data is available in Supplementary Material 07.

PQS in *Ascaris lumbricoides* genome. The ability of *A. lumbricoides* to maintain the highest relative level of stable PQS at high thresholds prompted us to analyze its G4 motifs in more details. To this aim, we selected the 2,313 sequences with a G4Hunter score above 2.0 (Supplementary Material 08) and analyzed them.

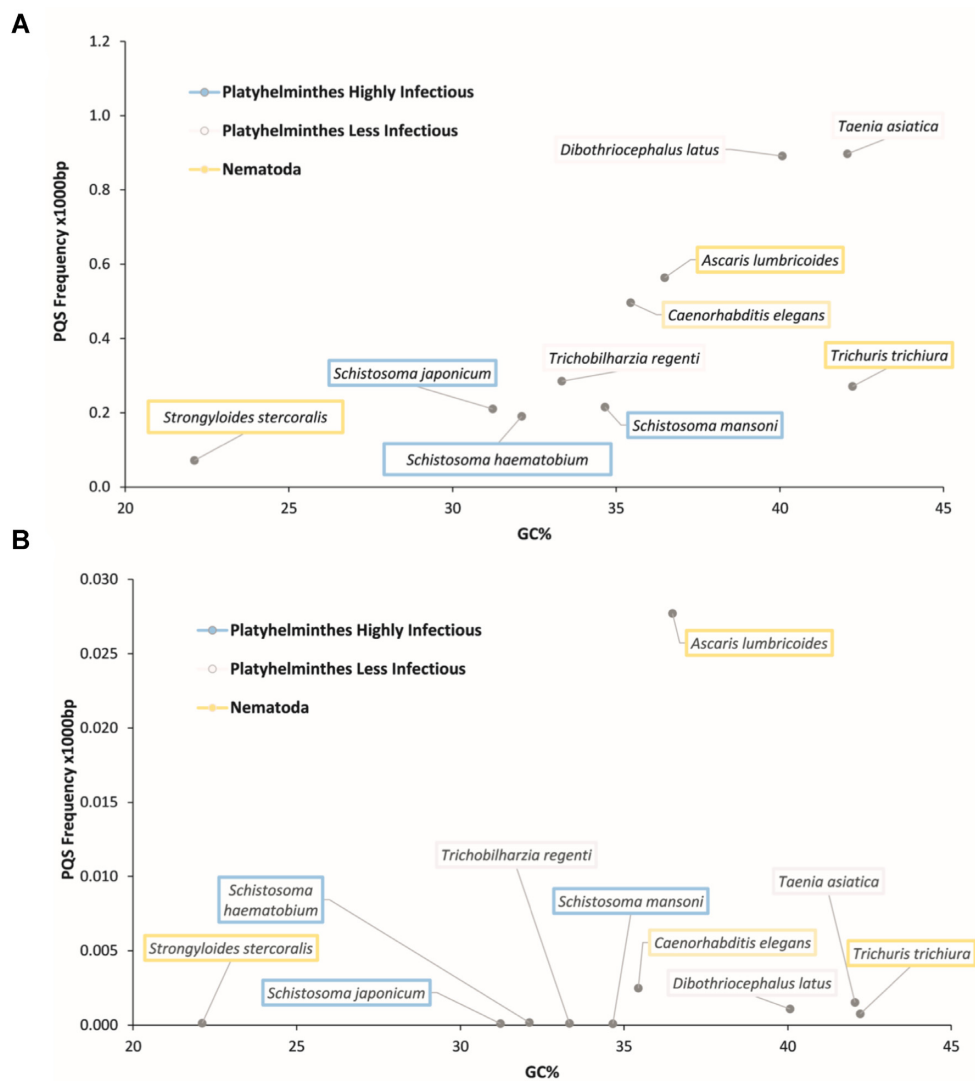


Figure 1. Relationship between GC% and PQS frequency per kb in genomic DNA (Color codes correspond to the groups (see Table 2, blue - Platyhelminths, yellow – Nematodes). Two different G4Hunter score thresholds were chosen for this analysis: (A) threshold ≥ 1.2 . (B) threshold ≥ 1.6 .

The most salient common feature of nearly all these *A. lumbricoides* motifs is that they are composed of poly-dG stretches, (or poly-dC when the score is negative) rather than other repetitive motifs (GGGT, GGGA, GGGGT, GGGGA, GGGGC, GGGGTT and GGGGAA would all give G4Hunter scores above 2.0). Out of 2,313 *A. lumbricoides* sequences with a G4Hunter score above 2.0 (or < 2.0), 2311 (99.9%) contain at least one run of at least 10 C/10 G. The length distribution of these runs is presented in Figure 3.

We then checked if a similar behavior was found for nematodes closely related to *A. lumbricoides* (40) (Supplementary Table S4). *T. canis* contained the highest number of PQS: 242,923 with a frequency of 0.77 PQS/kb and a GC content of 37.6% in the total genome. *T. canis* is followed by *A. suum*, with 182,227 PQS, with a frequency of 0.61 PQS/kb and a GC% of 37.7% (Supplementary Table S5). The majority of the PQS was found in the G4Hunter threshold 1.2–1.4 (476,476 PQS in total) with a frequency of 0.259 PQS/kb inside this range.

Programmed DNA elimination is a feature of nematodes such as *A. lumbricoides* and *A. suum*. This developmentally regulated process leads to the reproducible loss of specific genomic sequences in somatic cells, leaving the germline genome intact. In-depth analyses of DNA elimination in *A. lumbricoides* and *T. canis* have been performed recently (40). We analyzed G4 propensity in the breakpoint regions of *A. suum*, and found that 39 out of 40 regions contained at least one PQS nearby (within 3 kb), suggesting a possible role of G-quadruplexes in this process. When it comes to the total frequency comparison, the majority of the PQS found in *A. suum* are present before or inside the chromosomal break regions.

To test whether the PQS occurrence in chromosomal breakpoint regions is significantly different than in randomly shuffled sequences, we generated 40 random sequences with the same length and nucleotide content as original *A. suum* breakpoints. We found that 38 out of 40 randomly generated regions contained at least one PQS (compared to the original breakpoint regions where PQS

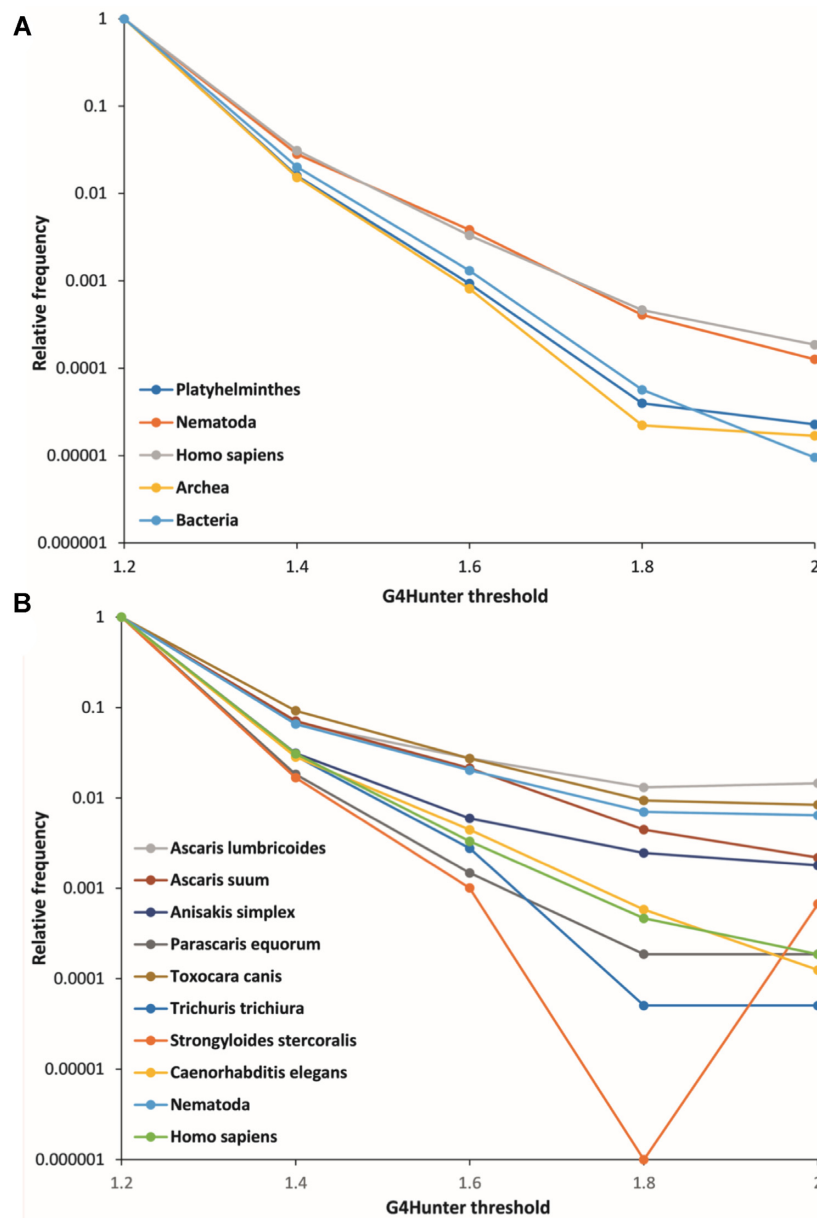


Figure 2. Distribution of G4 prone motifs per G4Hunter score: (A) in different phyla and (B) in selected species. *A. lumbricoides* exhibits an odd behavior (red curve) with a very high relative number of PQS with G4Hunter scores above 1.4). Data used to generate this panel can be found in Supplementary Table S3.

occurred in 39 out of 40 regions). However, the mean PQS count decreased from 9.25 PQS per breakpoint to 5.6 PQS. In addition, there was a highly significant difference between PQS counts ($P < 0.005$) in breakpoints and shuffled sequences. We performed a similar analysis with various G4Hunter thresholds (1.2–1.6). The higher the threshold, the higher the difference (and significance) between *A. suum* breakpoints and shuffled sequences. For a threshold of 1.6, 35 breakpoint sequences contain a least one PQS, while only 11 shuffled sequences contain one PQS (Figure 4; data provided in Sup Material 09).

To check if this trend was valid in other helminths, we performed additional analyzes of the breakpoints in two related species, *Toxocara canis* and *Parascaris equorum*,

thanks to the data collected by Wang *et al.* (44). We found a clear overrepresentation of G4 motifs in breakpoints as compared to shuffled sequences, and this result was valid at all threshold considered (see Supplementary Figure S7) as found in *A. suum*.

Experimental evidence for G4 formation *in vitro*

We identified a number of potential G4-forming motifs in helminth genomes using bioinformatics approaches. While G4Hunter's accuracy is reasonable, and actually excellent for high-scoring motifs (45), we found essential to confirm experimentally that some of these sequences are actually forming G4s, at least *in vitro*. To do so, we chose four se-

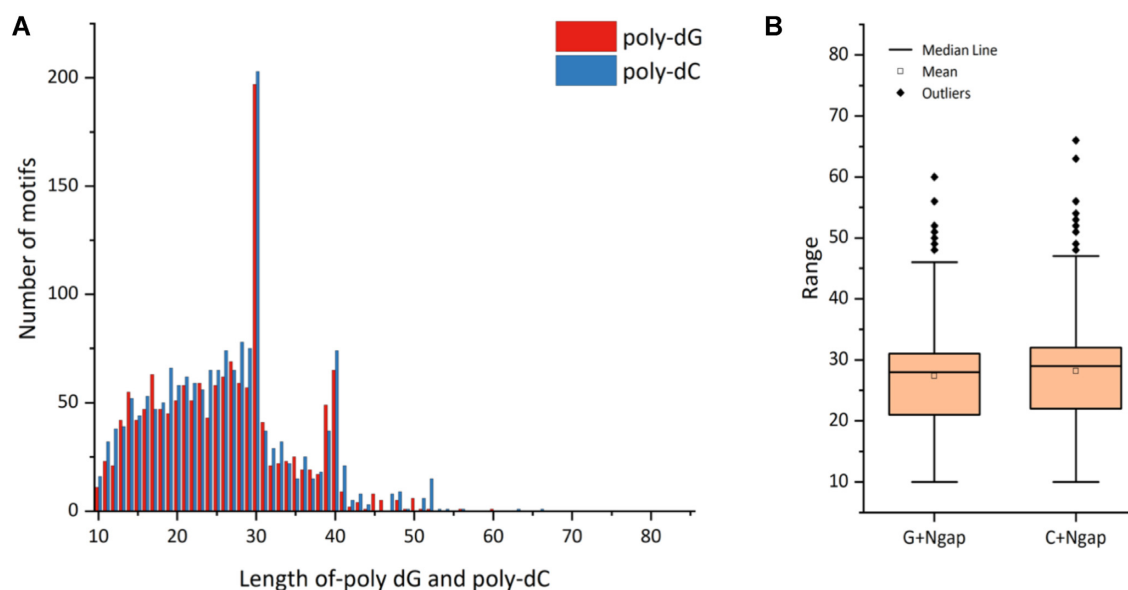


Figure 3. Distribution of pure homopolymeric G/C runs in *Ascaris lumbricoides* PQS motifs with a G4Hunter score above 2.0. (A) Length distribution for runs of 10 or more G (red) or C (blue). (B) Size distribution of G (left) and C (right) runs. Note that the genomic data include a number of N nucleotides corresponding to sequencing errors, adjacent to, or interrupting poly G/poly C runs. This is not very surprising, as reading error-prone sequences like these ones is difficult: this problem tends to underestimate the real length of poly dG stretches.

quences from *S. stercoralis*, three from *T. trichiura* as well as 20 sequences from *A. lumbricoides*. Results for *S. stercoralis* and *T. trichiura* are presented in Supplementary Information (Supplementary Table S6; Supplementary Figure S1-S2).

For the biophysical characterization of *A. lumbricoides* motifs, one should note that poly-dG runs have already been characterized in previous articles (46). For this reason, we focused our efforts on *A. lumbricoides* motifs that do not correspond to pure homopolymeric runs of guanines. These 20 sequences, which are 22–39 nucleotide-long, have G4Hunter scores between 1.46 and 2.48. G4 formation is considered extremely likely (>98.5%) for sequences with a score above 1.5 (45) but we wanted to provide compelling proof of G4 formation.

To demonstrate G4 formation, we used a combination of techniques, starting with FRET-MC, a method we very recently introduced (32) (Figure 5A). FRET-MC allows to test multiple sequences in parallel. A negative control (a sequence that does not form a G4 but a duplex, ds26) and two positive controls (sequences known to adopt G4 structures, Pu24T and c-myc) were used for comparison. The FRET-MC method measures the ability of a sequence to compete for binding to a well-known G4 ligand, PhenDC3. This compound is highly selective for G4s: efficient competitors are able to act as decoys for this G4 ligand, leading to a strong decrease in ΔT_m of a fluorescent G4-forming oligonucleotide (29); when added in large excess, these specific competitors can lead to negligible ΔT_m values. As can be seen in this panel, 19 out of 20 sequences considered here acted as efficient competitors (AL1 was found to be less efficient), arguing for G4 formation for most motifs.

Concluding on G4 formation based on a single technique is not recommended (45). Therefore, we used an independent approach, and investigated if the fluorescence emission

of G4 light-up probes (47), such as Thioflavin T (34) and NMM (48) was increased in the presence of *A. lumbricoides* motifs. Two negative and four positive controls were tested for comparison (Figure 5B, C). As shown in these panels, most sequences (including AL1 with NMM) induced significant increases in fluorescence emission to levels comparable or higher than the positive controls tested.

Finally, we performed additional spectroscopic experiments. G4s give specific signatures in isothermal and thermal difference spectra (IDS and TDS, respectively). The principle of these experiments is to compare the absorbance properties of the same oligonucleotide, in folded and unfolded state. The arithmetic difference between these two spectra gives a difference spectrum. Unfolding can be achieved by heating (for TDS) or by omitting stabilizing cations (for IDS). G4s exhibit a negative peak around 295 nm and a positive peak around 273 nm for both IDS and TDS (33). IDS and TDS of *Ascaris* motifs are shown in Supplementary Figure S3. Circular dichroism spectra of all AL sequences are presented in Supplementary Figure S4; interestingly, most spectra were indicative of parallel G4 formation (either exclusively or predominantly), except for AL14.

Altogether, these experiments confirmed predominant G4 formation for 19 out of the 20 *Ascaris* motifs tested. Of note, the motif for which a G4 may not be the dominant species and/or is of marginal stability (AL1) is the one with the lowest G4Hunter score (1.4).

Activity of G-quadruplex ligands against *Schistosoma mansoni*

We first verified that the *S. mansoni* G4 sequences recently described in (29) formed G4s, using the same techniques as the ones described for *A. lumbricoides*. As shown in Sup-

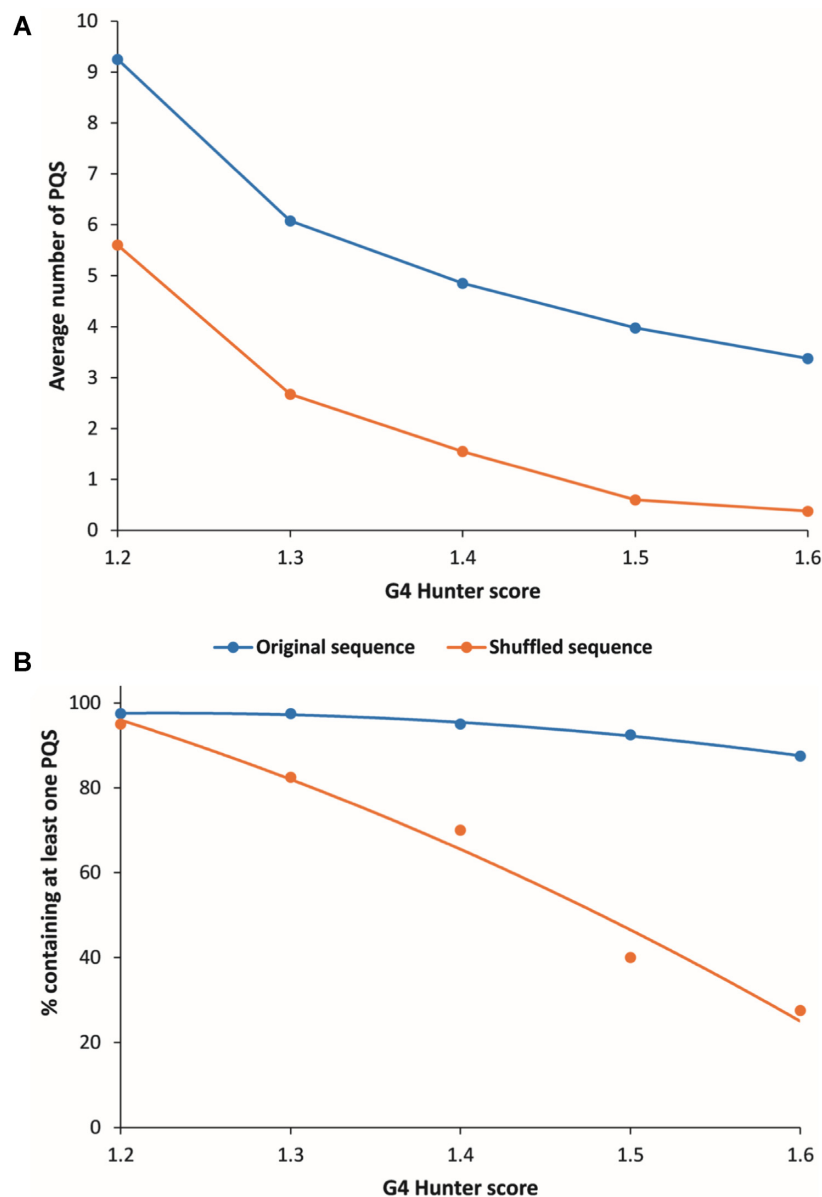


Figure 4. (A) Average number of PQS found in each breakpoint region, using different G4Hunter thresholds (x-axis) (B) fraction of breakpoint regions containing at least one quadruplex motif with a G4Hunter score above a certain threshold.

plementary Figure S5, this is indeed the case for the five positive sequences described in this article. Interestingly, for the sixth one (*smp-196840*), while the CD spectra reported by Craven *et al.* could not be associated with any known G4 CD profile, we were able to conclude that this motif was also able to form a stable G4, as shown by a combination of three independent methods (FRET-MC and two fluorescent light-up probes). The formation of a G4 by this motif is hardly surprising given its high G4Hunter score (2.0); *smp-196840* sequence is d-GGGAGGGGGAGAGA GAGAGGGGGAGGTAAGGG). Overall, we conclude that all six sequences investigated form stable G4s.

We next investigated whether G4 ligands (i.e. small compounds which selectively recognize this unusual nucleic acid structure) recently synthesized (Figure 6A) (35–36) would

bind to the *S. mansoni* G4s described in (29). Six compounds were chosen, with variable levels of stabilization of G4 structures. We performed a FRET-melting assay, in which we measured the melting temperature of a dual-labeled fluorescent G4-forming oligonucleotide (F21T, corresponding to the human telomeric motif, but also to *S. mansoni* telomeres). The tested compounds have variable affinities for the telomeric motif, with ΔT_m of +0 to +14°C. To verify that the active compounds were also able to recognize other *S. mansoni* quadruplexes, we added these sequences as unlabeled competitors. As shown in Figure 6B, the addition of some, but not all of these oligonucleotides led to a decrease in the stabilization induced by the G4 ligand considered here. This indicates that motifs such as *smp163240* or *smp319480* were able to act as efficient ‘de-

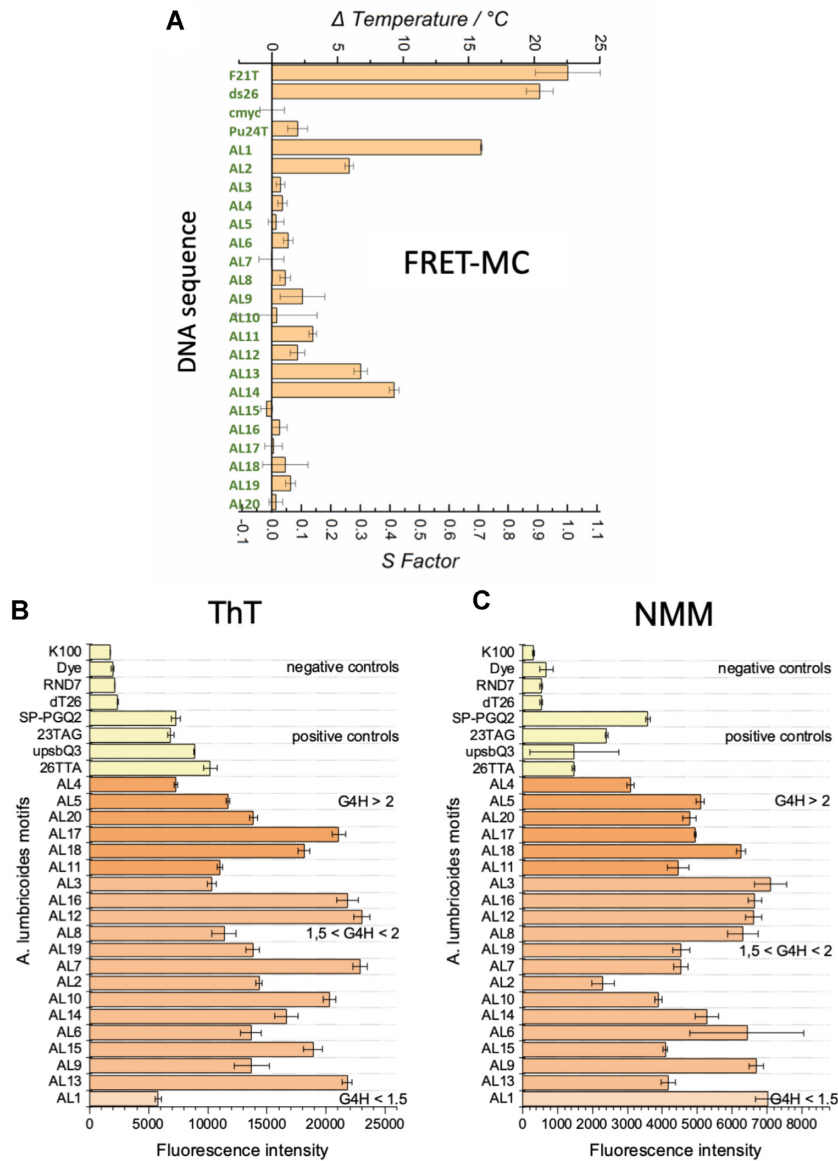


Figure 5. Evidence for G-quadruplex formation with 20 *Ascaris lumbricoides* sequences. (A) FRET MC results. (B) Thioflavin T fluorescence emission and (C) NMM fluorescence emission. Both Thioflavin T and NMM are light-up probes for which fluorescence intensity increases in the presence of G4-forming sequences.

coys' for the G4 ligands (even more efficient than the cmyc quadruplex used as a positive control), confirming that these molecules have an affinity for at least some of the G4 motifs (telomeric and non-telomeric) found in *S. mansoni* genome. In contrast, the ds26 negative control (double-stranded oligonucleotide) and two *S. mansoni* quadruplexes had little or no effect (Figure 6B), suggesting that JG1352 has little or no affinity for these structures. The FRET melting assay presented in Figure 7A illustrates that the best ligands (e.g. JG1352) stabilize all five G-quadruplexes tested in this assay to a variable extent, but do not stabilize the hairpin double-stranded control (FdxT; $\Delta T_m \approx 0$; sequences provided in Supplementary Table S7).

The next step was to determine if these G4 ligands would have an antiparasitic activity. We tested the biological activ-

ity of these six compounds against larval and adult *S. mansoni*. JG1057 and JG1352 showed high activity at 100 μ M and 10 μ M and moderate activity at 1 μ M against the larval stages. Both compounds revealed also high activity at 10 and 1 μ M against adult *S. mansoni*. JG966 revealed a lower activity in particular against adult *S. mansoni*, affecting adult *S. mansoni* only at the highest concentration of 10 μ M (Table 3). IC_{50} values calculated for JG1057 and JG1352 are in the range of praziquantel (a drug currently used to treat parasitic worm infections). Interestingly, the three inactive or weakly active compounds (ligands exhibiting low ΔT_m values on G-quadruplexes) (Figure 7A) were also significantly less active towards the parasite (Figure 7B) than two of the three best ligands. This effect was in particular against adult *S. mansoni* (both at 1 and 10 μ M

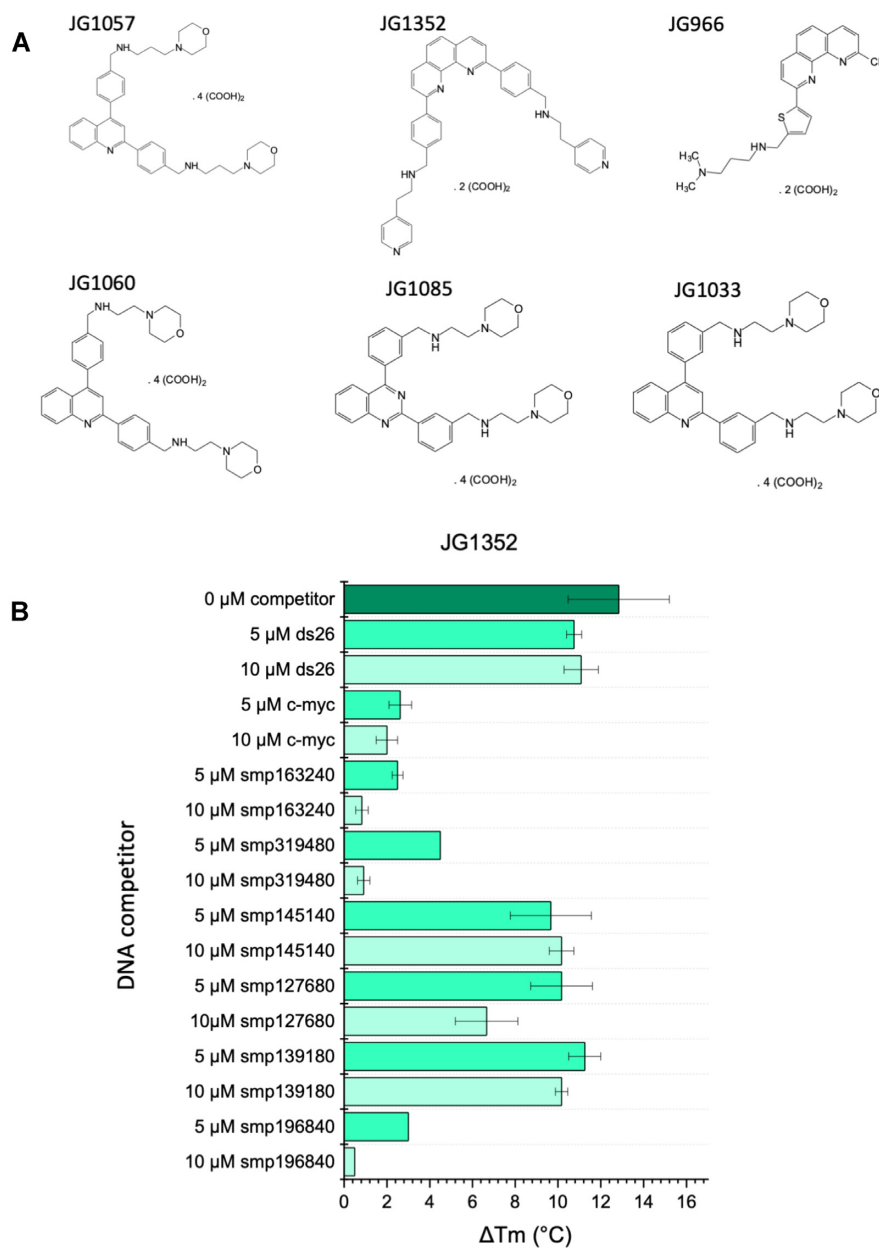


Figure 6. (A) Formula of the G4 ligands tested against *S. mansoni*. The three most active compounds (as determined by high ΔT_m values, see below) are shown on top; less active ligands are shown below. (B) Evidence that JG1352 binds to *S. mansoni* G-quadruplexes: FRET-melting data.

compound concentrations), suggesting that part of the antiparasitic effect of these compounds was mediated by a G4-related mechanism.

DISCUSSION

Besides the classical B-DNA double-helix structure, genomic DNA may adopt a variety of non-canonical structures which may play important roles (49). Repetitive sequences may form G4s (50) or i-DNA, inverted repeats can adopt cruciform structures (51), while CAG/CTG triplet repeats form unusual duplexes (52) and homopurine-homopyrimidine repeats adopt triplex structures (53). Many of these structures are involved in human patholo-

gies, such as neurological disorders or cancers (53–55). Recently their existence have been confirmed in living cells by several methods including secondary structures specific antibodies, synthetic compounds, and structure-sensitive sequencing (55–57). One of the most important local DNA structures seems to be G4, which has a better thermal stability compare to the B-DNA (58,59). Growing full-genome sequencing data provide an excellent source of information for detailed G4 prediction in various organisms. G4 has been shown to exist in archaea (43), bacteria (42,60,61) and eukaryote domains (19) as well in various viruses (62) where PQS propensity correlate with their host (63): viruses causing acute type of infections (including SARS-CoV2 genome) seem to be depleted for PQS (64,65). G4 have been

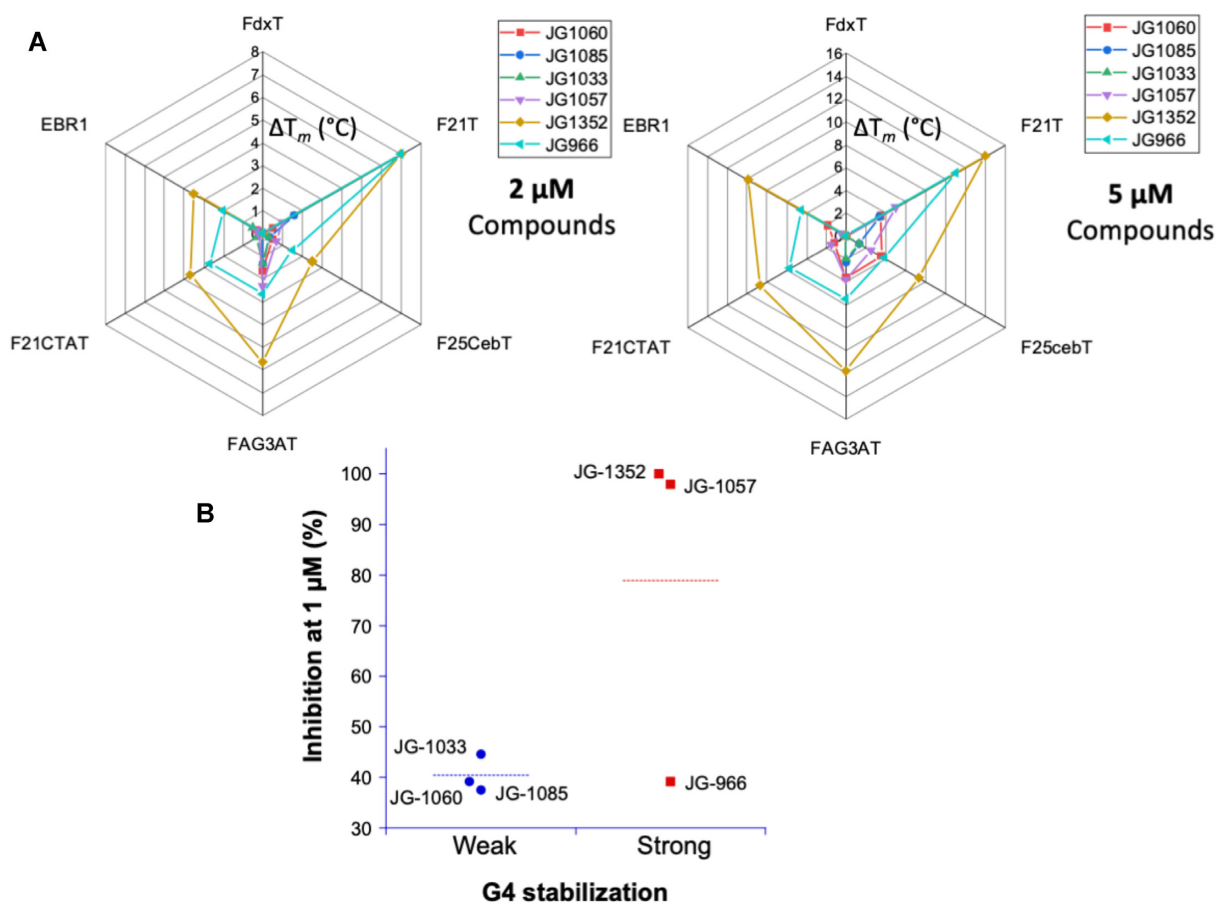


Figure 7. (A) ΔT_m values for the six compounds tested on six different nucleic acid structures, one duplex (FdxT) and five quadruplexes. Values are provided at two ligand concentrations, 2 (left) and 5 (right) μM . Sequences of the fluorescent oligonucleotides are provided in Supplementary Table S7. (B) Relation between activity against *S. mansoni* determined at 1 μM and G4 stabilization. Compounds were binned into two classes, 'weak' or 'strong', depending on ΔT_m on the human telomeric quadruplex (F21T; $\Delta T_m < 3.5^\circ\text{C}$ and $> 5^\circ\text{C}$, respectively).

Table 3. Antischistosomal activity of three G4 ligands tested at 3 different concentrations. Values correspond to growth inhibition (%). Compounds were tested in triplicate and *S. mansoni* incubated in the presence of the highest concentration of DMSO (<1%) served as control. The DMSO concentrations used in the assay (<1%) are routinely used in the lab, well tolerated and do not affect the parasite. Standard deviation is indicated within parentheses

Compound	Effect on newly transformed schistosomula (%) (SD)				Effect on adult <i>S. mansoni</i> (%) (SD)			
	100 μM	10 μM	1 μM	IC ₅₀ value (μM)	10 μM	1 μM	0.1 μM	IC ₅₀ value (μM)
JG 1057	100.0 (0)	92.5 (0.8)	46.7 (6.7)	1.49	98.2 (1.9)	97.9 (2.1)	29.6 (0)	0.11
JG 1352	100.0 (0)	100.0 (0)	56.7 (3.3)	0.4	100.0 (0)	100.0 (0)	31.5 (1.9)	0.07
JG 966	97.9 (2.1)	78.8 (7.9)	50.0 (10.0)	1.28	40.6 (0)	39.2 (0)	ND	>10
JG 1085	ND	100 (0)	32.7 (1.9)	1.18	48.2 (3.7)	37.4 (1.8)	ND	>10
JG 1060	ND	100 (0)	27.0 (3.8)	1.25	34.5 (3.6)	39.2 (3.6)	ND	>10
JG 1033	ND	100 (0)	25.0 (1.9)	1.28	48.2 (3.7)	44.6 (1.8)	ND	>10
Praziquantel ^a				2.2				0.1

^aReference for the IC₅₀ values of praziquantel taken from (80).

suggested to be valuable druggable targets for the development of a therapy against various pathogens (60,66). The recently published genome sequences of helminth species allowed us to perform comparative analyses of PQS in their genomes including broadly spread pathogenic species (4). We provided compelling evidence that many of these sequences form G4 structures *in vitro*. This does not necessarily imply that all of them adopt a G4 fold *in vivo*, as PQS were tested in the absence of flanking sequences, their reverse complement sequence, and proteins that could

promote or inhibit folding. Nevertheless, results previously published on *S. mansoni* with the BG4 antibody attest that at least some of the candidate motifs adopt a quadruplex fold *in vivo* (29).

Comparison of PQS in nuclear and mtDNA

Far less potential G4 motifs are found in mitochondrial DNA. This is due to the much smaller size of the mitochondrial genome, even when considering the number of PQS

per kb, a lower density is still found on mtDNA, and this observation is true both for platyhelminths and nematodes (Supplementary Figure S6). Overall, and in stark contrast to humans, where G4 are abundant in mtDNA, G4 seem relatively rare in helminth mtDNA. Differences in GC content partially account for these differences. Nuclear DNA and mtDNA of *S. mansoni*, *S. japonicum*, *S. haematobium* and *T. regenti* have relatively similar GC content, while for other helminths (*D. latus*, *T. asiatica*, *T. trichiura*, *S. stercoralis*, *A. lumbricoides*, *C. elegans*), the GC content of nuclear DNA is higher by 8% or more over mtDNA. Surprisingly, only one parasite (*T. regenti*) has a higher mitochondrial over nuclear GC content.

This relative paucity in mtDNA prevents an in-depth comparison of mt G4 density between helminths. Nevertheless, some motifs appear conserved between species and are located in gene regions with potential to regulate their expression. On the other hand, numerous PQS are found in the genomes of all ten helminths considered here. Unfortunately, due to lack of annotation in the most species, the only species for which a detailed analysis of the reference genome is possible is *C. elegans* (67,68). These analyses of PQS localization would be interesting to evidence putative G4 regulatory function in translation, which have already been demonstrated in other organisms (69–71).

Telomeric motifs

While Platyhelminths and humans share the same telomeric motif (TTAGGG)_n (72), Nematodes have a slightly different telomeric repeat (TTAGGC)_n (73). This one-nucleotide difference has a strong impact on G4 formation. The human telomeric DNA motif (TTAGGG)_n with a G4Hunter score of 1.5 is able to form a stable G4 both *in vitro* (74) and *in vivo* (75,76). Interestingly, both *S. haematobium* and *S. mansoni* have telomeric-like sequences integrated at non-telomeric sites (72). In contrast, the Nematode telomeric motif (TTAGGC)_n has a significantly lower G4Hunter score (0.5). Despite this low score, such motif may form an antiparallel G4 ($T_m \approx 40^\circ\text{C}$; Marquievillie, submitted), possibly in competition with a fundamentally different secondary structure called a foldback (77).

Programmed DNA elimination in nematodes

DNA elimination occurs in a number of species, including nematodes. It mostly corresponds to repetitive sequences and germline-specific genes. Previous analyses suggest that DNA elimination in nematodes silences germline-expressed genes (44). Their results suggested a sequence-independent mechanism for DNA breakage. Interestingly, we found that there is a clear overrepresentation of G4 motifs in break-points as compared to shuffled sequences, and this result was valid at all threshold considered (see Figure 4 and Supplementary Figure S7; Supplementary Material 9) and for three nematode species: *Ascaris suum*, *Toxocara canis* and *Parascaris equorum*. This conservation suggests that G-quadruplex formation may be involved in this programmed elimination, perhaps by recruiting a DNA cleaving complex (44). Telomere healing (that would lead to the insertion of a few telomeric motifs) would not explain this bias given the low G4Hunter score (see § above) of nematode telomeres.

Ascaris lumbricoides, a unique organism

Using G4Hunter score as a proxy for G4 stability, there is a strong counterselection against stable G4 in all Platyhelminths, in a manner similar to what is found in archaea and bacteria. G4 motifs with G4Hunter scores above 1.8 or 2.0 tend to be very rare, in comparison with the total number of motifs, as illustrated in Figure 2. This may suggest that Platyhelminths are unable to cope with very stable G4s, which cause problems during replication or transcription. On the other hand, nematodes follow a profile comparable to humans, where selection against stable G4 is not as strong as in prokaryotes.

Poly G runs frequency is worth discussing in the light of the analysis performed by Puig-Lombardi *et al.* (78). They analyzed the frequency of (GGGN)₃GGG motifs (15-nt sequence), corresponding to four runs of three guanines plus one extra nucleotide N. When N = G, this corresponds to ‘pure’ G₁₅ runs. This analysis was performed on over hundreds of genome assemblies. The authors found that the (GGGA)₃GGG motif was largely predominant in placental mammals, including humans, while an excess of (GGG)₃GGG (pure poly G runs) was found in amphibians, fish, plants, invertebrates and nematodes, including *C. elegans* (see Fig. S15 of reference (78)). The authors wrote that ‘the extreme prevalence (98%) of the G-runs versus the other GGGX motifs in the *C. elegans* genome and the finding that these sequences are eliminated by complete deletion during development and in animals deficient for the dog-1 helicase suggest that different molecular mechanisms can play a role in handling the equilibrium between the maintenance and the inactivation of short-loop G4-L1 motifs’.

What is unique about *A. lumbricoides* is therefore not the presence of pure poly G runs as compared to other repeats, but the overall abundancy of long PQS with high G4Hunter scores. *A. lumbricoides* and related species such as *A. suum* have an exceptional density of high stability G4s (with a high G4Hunter score); far more frequent than any other species. These results suggest that *A. lumbricoides* must have evolved very efficient mechanisms to cope with these stable secondary structures. Therefore, we checked if its genome contained putative helicases susceptible to unfold G4s. Dog-1 has been found to be involved in the genomic stability of poly dG stretches in *C. elegans*. For this reason, we looked for orthologs of the Dog-1 helicase domain (aa 80–440) in *A. lumbricoides*. Interestingly, four genes had homology, suggesting putative helicases involved in G4 resolution/unfolding (see Supplementary Table S8). Of note, one of them is an ortholog of a human ATP-dependent DNA helicase (also called DDX11, CHLR1 or KRG2) reported to act on G4 substrates (79). Further studies could evaluate the impact of G4 ligands on *Ascaris* spp.

CONCLUSION

Helminth genomes contain multiple PQS sequences. The comparison of various nematodes and Platyhelminths revealed interesting and marked differences between helminths. Experimental confirmation of G4 formation *in vitro* was obtained for four different species. Two of the G4 ligands able to bind to *Schistosoma mansoni* G4 motifs exhibited potent antiparasitic activity both against larval

and adult forms of this parasite, opening new perspectives for the use of G4 ligands to fight neglected tropical diseases. These results therefore open new perspectives for the development of novel therapeutic strategies against these widespread helminths.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank S. Amrane (IECB), L. Lacroix (ENS) and T. Mardivirin (LOB) for helpful discussions.

FUNDING

SYMBIT project [CZ.02.1.01/0.0/0.0/15_003/0000477] financed by the ERDF, INCa PL-Bio; ANR [ANR-20-CE12-0023 “G4Access”]; Inserm, CNRS and the Swiss National Science Foundation [320030_175585/1]. Funding for open access charge: SYMBIT.

Conflict of interest statement. None declared.

REFERENCES

- Hotez,P.J., Bundy,D.A.P., Beegle,K., Brooker,S., Drake,L., de Silva,N., Montresor,A., Engels,D., Jukes,M., Chitsulo,L. *et al.* (2006) Helminth Infections: soil-transmitted helminth infections and schistosomiasis. In: Jamison,D.T., Breman,J.G., Measham,A.R., Alleyne,G., Claeson,M., Evans,D.B., Jha,P., Mills,A. and Musgrove,P. (eds). *Disease Control Priorities in Developing Countries*. 2nd edn. Washington (DC).
- Bethony,J., Brooker,S., Albonico,M., Geiger,S.M., Loukas,A., Diemert,D. and Hotez,P.J. (2006) Soil-transmitted helminth infections: ascariasis, trichuriasis and hookworm. *The Lancet*, **367**, 1521–1532.
- Verjee,MA. (2019) Schistosomiasis: still a cause of significant morbidity and mortality. *Res. Rep. Trop. Med.*, **10**, 153–163.
- Anderson,T.J.C. and Duraisingh,M.T. (2020) Transformative tools for parasitic flatworms. *Science*, **369**, 1562–1564.
- Christian,P., Khatry,S.K. and West,K.P. (2004). Antenatal anthelmintic treatment, birthweight, and infant survival in rural Nepal. *Lancet*, **364**, 981–983.
- Charlier,J., De Waele,V., Ducheyne,E., van der Voort,M., Vande Velde,F. and Claerebout,E. (2016) Decision making on helminths in cattle: diagnostics, economics and human behaviour. *Irish Vet. J.*, **69**, 14.
- de Silva,N.R., Brooker,S., Hotez,P.J., Montresor,A., Engels,D. and Savioli,L. (2003) Soil-transmitted helminth infections: updating the global picture. *Trends Parasitol.*, **19**, 547–551.
- Nutman,T.B. (2017) Human infection with *Strongyloides stercoralis* and other related *Strongyloides* species. *Parasitology*, **144**, 263–273.
- Colley,D.G., Bustinduy,A.L., Secor,W.E. and King,C.H. (2014) Human schistosomiasis. *Lancet*, **383**, 2253–2264.
- Thétiot-Laurent,S.A.-L., Boissier,J., Robert,A. and Meunier,B. (2013) Schistosomiasis chemotherapy. *Angew. Chem. Int. Ed.*, **52**, 7936–7956.
- Furtado,L.F.V., de Paiva Bello,A.C.P. and Rabelo,É.M.L. (2016) Benzimidazole resistance in helminths: from problem to diagnosis. *Acta Tropica*, **162**, 95–102.
- Bergquist,R., Zhou,X.-N., Rollinson,D., Reinhard-Rupp,J. and Klohe,K. (2017) Elimination of schistosomiasis: the tools required. *Infect. Dis. Poverty*, **6**, 158.
- Berriman,M., Haas,B.J., LoVerde,P.T., Wilson,R.A., Dillon,G.P., Cerqueira,G.C., Mashiyama,S.T., Al-Lazikani,B., Andrade,L.F., Ashton,P.D. *et al.* (2009) The genome of the blood fluke *Schistosoma mansoni*. *Nature*, **460**, 352–358.
- Zhou,Y., Zheng,H., Chen,Y., Zhang,L., Wang,K., Guo,J., Huang,Z., Zhang,B., Huang,W., Jin,K. *et al.* (2009) The *Schistosoma japonicum* genome reveals features of host–parasite interplay. *Nature*, **460**, 345–351.
- Young,N.D., Jex,A.R., Li,B., Liu,S., Yang,L., Xiong,Z., Li,Y., Cantacessi,C., Hall,R.S., Xu,X. *et al.* (2012) Whole-genome sequence of *Schistosoma haematobium*. *Nat. Genet.*, **44**, 221–225.
- Kolesnikova,S. and Curtis,E.A. (2019) Structure and function of multimeric G-quadruplexes. *Molecules*, **24**, 3074.
- Burge,S., Parkinson,G.N., Hazel,P., Todd,A.K. and Neidle,S. (2006) Quadruplex DNA: sequence, topology and structure. *Nucleic Acids Res.*, **34**, 5402–5415.
- Lane,A.N., Chaires,J.B., Gray,R.D. and Trent,J.O. (2008) Stability and kinetics of G-quadruplex structures. *Nucleic Acids Res.*, **36**, 5482–5515.
- Yoshida,W., Saikyo,H., Nakabayashi,K., Yoshioka,H., Bay,D.H., Iida,K., Kawai,T., Hata,K., Ikebukuro,K., Nagasawa,K. *et al.* (2018) Identification of G-quadruplex clusters by high-throughput sequencing of whole-genome amplified products with a G-quadruplex ligand. *Sci. Rep.*, **8**, 3116.
- Carvalho,J., Mergny,J.L., Salgado,G.F., Queiroz,J.A. and Cruz,C. (2020) G-quadruplex, friend or foe: the role of the G-quartet in anticancer strategies. *Trends Mol. Med.*, **26**, 848–861.
- Balasubramanian,S., Hurley,L.H. and Neidle,S. (2011) Targeting G-quadruplexes in gene promoters: a novel anticancer strategy? *Nat. Rev. Drug Discov.*, **10**, 261–275.
- Siddiqui-Jain,A., Grand,C.L., Bearss,D.J. and Hurley,L.H. (2002) Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress c-MYC transcription. *Proc. Natl. Acad. Sci. U.S.A.*, **99**, 11593–11598.
- Sun,Z.-Y., Wang,X.-N., Cheng,S.-Q., Su,X.-X. and Ou,T.-M. (2019) Developing novel G-quadruplex ligands: from interaction with nucleic acids to interfering with nucleic acid–protein interaction. *Molecules*, **24**, 396.
- Zhao,J., Bacolla,A., Wang,G. and Vasquez,K.M. (2010) Non-B DNA structure-induced genetic instability and evolution. *Cell. Mol. Life Sci.*, **67**, 43–62.
- Wong,H.M. and Huppert,J.L. (2009). Stable G-quadruplexes are found outside nucleosome-bound regions. *Mol. BioSyst.*, **5**, 1713–1719.
- Marsico,G., Chambers,V.S., Sahakyan,A.B., McCauley,P., Boutell,J.M., di Antonio,M. and Balasubramanian,S. (2019) Whole genome experimental maps of DNA G-quadruplexes in multiple species. *Nucleic Acids Res.*, **47**, 3862–3874.
- Hoffmann,R.F., Moshkin,Y.M., Mouton,S., Grzeschik,N.A., Kalicharan,R.D., Kuipers,J., Wolters,A.H.G., Nishida,K., Romashchenko,A.V., Postberg,J. *et al.* (2016) Guanine quadruplex structures localize to heterochromatin. *Nucleic Acids Res.*, **44**, 152–163.
- Yu,Q.-Q., Gao,J.-J., Lang,X.-X., Li,H.-Y. and Wang,M.-Q. (2021) Microenvironment-sensitive fluorescent ligand binds ascariis telomere antiparallel G-quadruplex DNA with blue-shift and enhanced emission. *ChemBioChem.*, **22**, 1042–1048.
- Craven,H.M., Bonsignore,R., Lenis,V., Santi,N., Berrar,D., Swain,M., Whiteland,H., Casini,A. and Hoffmann,K.F. (2021) Identifying and validating the presence of guanine-quadruplexes (G4) within the blood fluke parasite *Schistosoma mansoni*. *PLoS Negl. Trop. Dis.*, **15**, e0008770.
- Brázda,V., Kolomazník,J., Lýsek,J., Bartas,M., Fojta,M., Štastný,J. and Mergny,J.L. (2019) G4Hunter web application: a web server for G-quadruplex prediction. *Bioinformatics*, **35**, 3493–3495.
- Mergny,J.L., Phan,A.T. and Lacroix,L. (1998) Following G-quartet formation by UV-spectroscopy. *FEBS Lett.*, **435**, 74–78.
- Luo,Y., Granzhan,A., Verga,D. and Mergny,J.-L. (2021) FRET-MC: a fluorescence melting competition assay for studying G4 structures in vitro. *Biopolymers*, **112**, e23415.
- Mergny,J.-L., Li,J., Lacroix,L., Amrane,S. and Chaires,J.B. (2005) Thermal difference spectra: a specific signature for nucleic acid structures. *Nucleic Acids Res.*, **33**, e138.
- Renaud de la Faverie,A., Guédin,A., Bedrat,A., Yatsunyk,L.A. and Mergny,J.-L. (2014) Thioflavin T as a fluorescence light-up probe for G4 formation. *Nucleic Acids Res.*, **42**, e65.
- Guillon,J., Denevault-Sabourin,C., Chevret,E., Brachet-Botineau,M., Milano,V., Guédin-Beaupaire,A., Moreau,S., Ronga,L.,

- Savrimoutou, S., Rubio, S. *et al.* (2021) Design, synthesis, and antiproliferative effect of 2,9-bis[4-(pyridinylalkylaminomethyl)phenyl]-1,10-phenanthroline derivatives on human leukemic cells by targeting G-quadruplex. *Archiv. Pharmazie*, **354**, e2000450.
36. Guillon, J., Cohen, A., Das, R.N., Boudot, C., Gueddouda, N.M., Moreau, S., Ronga, L., Savrimoutou, S., Basmaciyan, L., Tisnerat, C. *et al.* (2018) Design, synthesis, and antiprotozoal evaluation of new 2,9-bis[(substituted-aminomethyl)phenyl]-1,10-phenanthroline derivatives. *Chem. Biol. Drug Des.*, **91**, 974–995.
 37. Guillon, J., Cohen, A., Boudot, C., Valle, A., Milano, V., Das, R.N., Guédin, A., Moreau, S., Ronga, L., Savrimoutou, S. *et al.* (2020) Design, synthesis, and antiprotozoal evaluation of new 2,4-bis[(substituted-aminomethyl)phenyl]isoquinoline, 1,3-bis[(substituted-aminomethyl)phenyl]isoquinoline and 2,4-bis[(substituted-aminomethyl)phenyl]quinazoline derivatives. *J. Enz. Inhib. Med. Chem.*, **35**, 432–459.
 38. Lombardo, F.C., Pasche, V., Panic, G., Endriss, Y. and Keiser, J. (2019) Life cycle maintenance and drug-sensitivity assays for early drug discovery in *Schistosoma mansoni*. *Nat. Protoc.*, **14**, 461–481.
 39. Crooks, G.E., Hon, G., Chandonia, J.-M. and Brenner, S.E. (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.
 40. Coghlan, A., Tyagi, R., Cotton, J.A., Holroyd, N., Rosa, B.A., Tsai, I.J., Laetsch, D.R., Beech, R.N., Day, T.A., Hallsworth-Pepin, K. *et al.* (2019) Comparative genomics of the major parasitic worms. *Nat. Gen.*, **51**, 163–174.
 41. Jourdan, P.M., Lambertson, P.H.L., Fenwick, A. and Addiss, D.G. (2018) Soil-transmitted helminth infections. *Lancet*, **391**, 252–265.
 42. Bartas, M., Čutová, M., Brázda, V., Kaura, P., Štátný, J., Kolomazník, J., Coufal, J., Goswami, P., Červen, J. and Pečinka, P. (2019) The presence and localization of G-quadruplex forming sequences in the domain of bacteria. *Molecules*, **24**, 1711.
 43. Brázda, V., Luo, Y., Bartas, M., Kaura, P., Porubiaková, O., Štátný, J., Pečinka, P., Verga, D., Da Cunha, V., Takahashi, T.S. *et al.* (2020) G-Quadruplexes in the archaea domain. *Biomolecules*, **10**, 1349.
 44. Wang, J., Gao, S., Mostovoy, Y., Kang, Y., Zagoskin, M., Sun, Y., Zhang, B., White, L.K., Easton, A., Nutman, T.B. *et al.* (2017) Comparative genome analysis of programmed DNA elimination in nematodes. *Genome Res.*, **27**, 2001–2014.
 45. Bedrat, A., Lacroix, L. and Mergny, J.-L. (2016) Re-evaluation of G-quadruplex propensity with G4Hunter. *Nucleic Acids Res.*, **44**, 1746–1759.
 46. Sengar, A., Heddi, B. and Phan, A.T. (2014) Formation of G-quadruplexes in poly-G sequences: structure of a propeller-type parallel-stranded G-quadruplex formed by a G15 stretch. *Biochemistry*, **53**, 7718–7723.
 47. Largy, E., Granzhan, A., Hamon, F., Verga, D. and Teulade-Fichou, M.P. (2013) Visualizing the quadruplex: from fluorescent ligands to light-up probes. *Top. Curr. Chem.*, **330**, 111–177.
 48. Sabharwal, N.C., Savikhin, V., Turek-Herman, J.R., Nicoludis, J.M., Szalai, V.A. and Yatsunyk, L.A. (2014) N-methylmesoporphyrin IX fluorescence as a reporter of strand orientation in guanine quadruplexes. *FEBS J.*, **281**, 1726–1737.
 49. McKinney, J.A., Wang, G., Mukherjee, A., Christensen, L., Subramanian, S.H.S., Zhao, J. and Vasquez, K.M. (2020) Distinct DNA repair pathways cause genomic instability at alternative DNA structures. *Nat Commun.*, **11**, 236.
 50. Spiegel, J., Adhikari, S. and Balasubramanian, S. (2020) The structure and function of DNA G-quadruplexes. *Trends Chem.*, **2**, 123–136.
 51. Brázda, V., Laister, R.C., Jagelská, E.B. and Arrowsmith, C. (2011) Cruciform structures are a common DNA feature important for regulating biological processes. *BMC Mol Biol.*, **12**, 33.
 52. Petruska, J., Arnheim, N. and Goodman, M.F. (1996) Stability of intrastrand hairpin structures formed by the CAG/CTG class of DNA triplet repeats associated with neurological diseases. *Nucleic Acids Res.*, **24**, 1992–1998.
 53. Helma, R., Bažantová, P., Petr, M., Adámik, M., Renčíuk, D., Tichý, V., Pastuchová, A., Soldánová, Z., Pečinka, P., Bowater, R.P. *et al.* (2019) p53 binds preferentially to non-B DNA structures formed by the pyrimidine-rich strands of GAA.TTC trinucleotide repeats associated with Friedreich's ataxia. *Molecules*, **24**, 2078.
 54. Cimino-Reale, G., Zaffaroni, N. and Folini, M. (2016) Emerging role of G-quadruplex DNA as target in anticancer therapy. *Curr. Pharm. Des.*, **22**, 6612–6624.
 55. Hänsel-Hertsch, R., Simeone, A., Shea, A., Hui, W.W.I., Zyner, K.G., Marsico, G., Rueda, O.M., Bruna, A., Martin, A., Zhang, X. *et al.* (2020) Landscape of G-quadruplex DNA structural regions in breast cancer. *Nat. Genet.*, **52**, 878–883.
 56. Poggi, L. and Richard, G.-F. (2021) Alternative DNA structures in vivo: molecular evidence and remaining questions. *Microbiol. Mol. Biol. Rev.*, **85**, e00110-20.
 57. Yang, C., Hu, R., Li, Q., Li, S., Xiang, J., Guo, X., Wang, S., Zen, Y. and Yang, G. (2018) Visualization of parallel G-quadruplexes in cells with a series of new developed Bis(4-aminobenzylidene)acetone derivatives. *ACS Omega*, **3**, 10487–10492.
 58. Bochman, M.L., Paeschke, K. and Zakian, V.A. (2012) DNA secondary structures: stability and function of G-quadruplex structures. *Nat. Rev. Genet.*, **13**, 770–780.
 59. Takahashi, S. and Sugimoto, N. (2013) Effect of pressure on thermal stability of G-quadruplex DNA and double-stranded DNA structures. *Molecules*, **18**, 13297–13319.
 60. Yadav, P., Kim, N., Kumari, M., Verma, S., Sharma, T.K., Yadav, V. and Kumar, A. (2021) G-Quadruplex structures in bacteria - biological relevance and potential as antimicrobial target. *J. Bacteriol.*, **203**, e0057720.
 61. Brázda, V., Fojta, M. and Bowater, R.P. (2020) Structures and stability of simple DNA repeats from bacteria. *Biochem J.*, **477**, 325–339.
 62. Ruggiero, E. and Richter, S.N. (2020) Viral G-quadruplexes: new frontiers in virus pathogenesis and antiviral therapy. *Annu. Rep. Med. Chem.*, **54**, 101–131.
 63. Bohálová, N., Cantara, A., Bartas, M., Kaura, P., Štátný, J., Pečinka, P., Fojta, M. and Brázda, V. (2021) Tracing dsDNA virus-host coevolution through correlation of their G-quadruplex-forming sequences. *Int. J. Mol. Sci.*, **22**, 3433.
 64. Bohálová, N., Cantara, A., Bartas, M., Kaura, P., Štátný, J., Pečinka, P., Fojta, M., Mergny, J.L. and Brázda, V. (2021) Analyses of viral genomes for G-quadruplex forming sequences reveal their correlation with the type of infection. *Biochimie*, **186**, 13–27.
 65. Bartas, M., Brázda, V., Bohálová, N., Cantara, A., Volná, A., Stachurová, T., Malachová, K., Jagelská, E.B., Porubiaková, O., Červeň, J. *et al.* (2020) In-depth bioinformatic analyses of human SARS-CoV-2, SARS-CoV, MERS-CoV, and other nidovirales suggest important roles of noncanonical nucleic acid structures in their lifecycles. *Front. Microbiol.*, **11**, 1583.
 66. Ruggiero, E. and Richter, S.N. (2020) Viral G-quadruplexes: new frontiers in virus pathogenesis and antiviral therapy. *Annu. Rep. Med. Chem.*, **54**, 101–131.
 67. Li, R., Hsieh, C.-L., Young, A., Zhang, Z., Ren, X. and Zhao, Z. (2015) Illumina synthetic long read sequencing allows recovery of missing sequences even in the «Finished» *C. elegans* genome. *Sci Rep.*, **5**, 10814.
 68. Kruijselbrink, E., Guryev, V., Brouwer, K., Pontier, D.B., Cuppen, E. and Tijsterman, M. (2008) Mutagenic capacity of endogenous G4 DNA underlies genome instability in FANCD1-defective *C. elegans*. *Curr Biol.*, **18**, 900–905.
 69. Vannutelli, A., Belhamiti, S., Garant, J.-M., Ouangraoua, A. and Perreault, J.-P. (2020) Where are G-quadruplexes located in the human transcriptome? *NAR Genom. Bioinform.*, **2**, lqaa035.
 70. Maltby, C.J., Schofield, J.P.R., Houghton, S.D., O'Kelly, I., Vargas-Caballero, M., Deinhardt, K. and Coldwell, M.J. (2020) A 5' UTR GGN repeat controls localisation and translation of a potassium leak channel mRNA through G-quadruplex formation. *Nucleic Acids Res.*, **48**, 9822–9839.
 71. Katsuda, Y., Sato, S.-I., Asano, L., Morimura, Y., Furuta, T., Sugiyama, H., Hagihara, M. and Uesugi, M. (2016) A small molecule that represses translation of G-quadruplex-containing mRNA. *J. Am. Chem. Soc.*, **138**, 9037–9040.
 72. Hirai, H. (2014) Chromosomal differentiation of schistosomes: what is the message? *Front. Genet.*, **5**, 301.
 73. Müller, F., Wicky, C., Spicher, A. and Tobler, H. (1991) New telomere formation after developmentally regulated chromosomal breakage during the process of chromatin diminution in *Ascaris lumbricoides*. *Cell*, **67**, 815–822.
 74. Phan, A.T. (2010) Human telomeric G-quadruplex: structures of DNA and RNA sequences. *FEBS J.*, **277**, 1107–1117.

75. Hänsel,R., Foldynová-Trantírková,S., Löhr,F., Buck,J., Bongartz,E., Bamberg,E., Schwalbe,H., Dötsch,V. and Trantírek,L. (2009) Evaluation of parameters critical for observing nucleic acids inside living *Xenopus laevis* oocytes by in-cell NMR spectroscopy. *J. Am. Chem. Soc.*, **131**, 15761–15768.
76. Biffi,G., Tannahill,D., McCafferty,J. and Balasubramanian,S. (2013) Quantitative visualization of DNA G-quadruplex structures in human cells. *Nat. Chem.*, **5**, 182–186.
77. Školáková,P., Foldynová-Trantírková,S., Bednářová,K., Fiala,R., Vorlíčková,M. and Trantírek,L. (2015) Unique *C. elegans* telomeric overhang structures reveal the evolutionarily conserved properties of telomeric DNA. *Nucleic Acids Res.* **43**, 4733–4745.
78. Puig Lombardi,E., Holmes,A., Verga,D., Teulade-Fichou,M.-P., Nicolas,A. and Londoño-Vallejo,A. (2019) Thermodynamically stable and genetically unstable G-quadruplexes are depleted in genomes across species. *Nucleic Acids Res.*, **47**, 6098–6113.
79. Ribeiro de Almeida,C., Dhir,S., Dhir,A., Moghaddam,A.E., Sattentau,Q., Meinhart,A. and Proudfoot,N.J. (2018) RNA helicase DDX1 converts RNA G-quadruplex structures into R-loops to promote IgH class switch recombination. *Mol. Cell.*, **70**, 650–662.
80. Ingram,K., Yaremenko,I.A., Krylov,I.B., Hofer,L., Terent'ev,A.O. and Keiser,J. (2012) Identification of antischistosomal leads by evaluating bridged 1,2,4,5-tetraoxanes, alphaperoxides, and tricyclic monoperoxides. *J. Med. Chem.*, **55**, 8700–8711.



Article

The Newly Sequenced Genome of *Pisum sativum* Is Replete with Potential G-Quadruplex-Forming Sequences—Implications for Evolution and Biological Regulation

Michaela Dobrovolná^{1,2}, Natália Bohálová^{1,3} , Vratislav Peška¹ , Jiawei Wang⁴, Yu Luo^{4,5} , Martin Bartas⁶ , Adriana Volná⁷, Jean-Louis Mergny^{1,4,*} and Václav Brázda^{1,2,*}

- ¹ Institute of Biophysics of the Czech Academy of Sciences, 612 65 Brno, Czech Republic; dobrovolna@ibp.cz (M.D.); natalia.bohalova@ibp.cz (N.B.); vpeska@ibp.cz (V.P.)
- ² Faculty of Chemistry, Brno University of Technology, Purkyňova 118, 612 00 Brno, Czech Republic
- ³ Department of Experimental Biology, Faculty of Science, Masaryk University, 611 37 Brno, Czech Republic
- ⁴ Laboratoire d'Optique et Biosciences (LOB), Ecole Polytechnique, CNRS, INSERM, Institut Polytechnique de Paris, CEDEX, 91128 Palaiseau, France; jiawei.wang@polytechnique.edu (J.W.); yu.luo@curie.fr (Y.L.)
- ⁵ CNRS UMR9187, INSERM U1196, Université Paris-Saclay, CEDEX, 91405 Orsay, France
- ⁶ Department of Biology and Ecology, Faculty of Science, University of Ostrava, 710 00 Ostrava, Czech Republic; martin.bartas@osu.cz
- ⁷ Department of Physics, Faculty of Science, University of Ostrava, 710 00 Ostrava, Czech Republic; adriana.volna@osu.cz
- * Correspondence: jean-louis.mergny@polytechnique.edu (J.-L.M.); vaclav@ibp.cz (V.B.)



Citation: Dobrovolná, M.; Bohálová, N.; Peška, V.; Wang, J.; Luo, Y.; Bartas, M.; Volná, A.; Mergny, J.-L.; Brázda, V. The Newly Sequenced Genome of *Pisum sativum* Is Replete with Potential G-Quadruplex-Forming Sequences—Implications for Evolution and Biological Regulation. *Int. J. Mol. Sci.* **2022**, *23*, 8482. <https://doi.org/10.3390/ijms23158482>

Academic Editor: Zsófia Bánfalvi

Received: 6 July 2022

Accepted: 28 July 2022

Published: 30 July 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Abstract: G-quadruplexes (G4s) have been long considered rare and physiologically unimportant in vitro curiosities, but recent methodological advances have proved their presence and functions in vivo. Moreover, in addition to their functional relevance in bacteria and animals, including humans, their importance has been recently demonstrated in evolutionarily distinct plant species. In this study, we analyzed the genome of *Pisum sativum* (garden pea, or the so-called green pea), a unique member of the *Fabaceae* family. Our results showed that this genome contained putative G4 sequences (PQSs). Interestingly, these PQSs were located nonrandomly in the nuclear genome. We also found PQSs in mitochondrial (mt) and chloroplast (cp) DNA, and we experimentally confirmed G4 formation for sequences found in these two organelles. The frequency of PQSs for nuclear DNA was 0.42 PQSs per thousand base pairs (kbp), in the same range as for cpDNA (0.53/kbp), but significantly lower than what was found for mitochondrial DNA (1.58/kbp). In the nuclear genome, PQSs were mainly associated with regulatory regions, including 5'UTRs, and upstream of the rRNA region. In contrast to genomic DNA, PQSs were located around RNA genes in cpDNA and mtDNA. Interestingly, PQSs were also associated with specific transposable elements such as TIR and LTR and around them, pointing to their role in their spreading in nuclear DNA. The nonrandom localization of PQSs uncovered their evolutionary and functional significance in the *Pisum sativum* genome.

Keywords: G-quadruplex; G4 propensity; chloroplast DNA; sequence prediction

1. Introduction

Pisum sativum, commonly known as the garden pea or green pea, is an important and broadly cultivated crop worldwide. It was domesticated ~10,000 years ago in the Near East [1]. Its seeds are rich in proteins, fibers, vitamins, minerals, and antioxidants [2]. In addition, the pea is widely used as a model plant species nowadays [3,4], and also is a historically important genetic model as the first organism for which the basic genetics laws were described and demonstrated by the Moravian monk Gregor Johann Mendel in 1865 [5,6]. His systematic work, statistic evaluation, and mathematical descriptions of his experiments with hereditary of seven independent pea features paved the foundation of

modern genetics. His discoveries were later called the laws of Mendelian inheritance in his honor.

G-quadruplexes (G4s) are four-stranded DNA or RNA structures in which alternative Hoogsteen base pairing (G-G) enables guanine tetrad formation. Each guanine tetrad corresponds to one stack of G4 structure and is stabilized by an internal spine of positively charged ions, mostly sodium (Na^+) or potassium (K^+). Depending on the number of guanine tetrads (stacks) we can distinguish two-, three-, four-, five-, or even six-stacked G4s [7]. As described above, G4 formation requires guanines—at least eight for a two-tetrad structure—and is thus favored in regions locally enriched in this nucleotide. From a functional perspective, G4s have been documented to influence replication, transcription, and even translation, which illustrates their importance in basic physiological cellular processes and indicates the need for their precise regulation [8–10]. A recent study showed the distinct roles of G4 in the transcription regulation of the rice genome based on its genomic localization. G4 found in promoters had a potentiated effect, whereas gene location caused repression of gene transcription [11].

The presence of G4s has been demonstrated in viral [12,13], bacterial [14], archaeal [15], fungal [16,17], and other eukaryotic genomes, including that of humans [18]. However, only a few genome-wide analyses of G4s in plants have been reported [11,19–22]. Genome-wide analyses of G4s have not been reported for *P. sativum*, the genome of which was sequenced and assembled 3 years ago [23]. *P. sativum* nuclear genome is composed of two metacentric and five acrocentric chromosomes [24]. The pea genome is relatively large (4.45 Gb) compared to other *Fabaceae*, such as *Glycine max* (soybean)—995 Mb, *Medicago truncatula* (barrel medic)—(420 Mb, or *Lotus japonicus* (bird's-foot trefoil)—385 Mb, mostly due to genome expansion of transposons, which comprise about 76% of the pea genome [25]. Several analysts suggested faster evolution of the pea genome in comparison with the species mentioned above due to frequent recombination events mediated by transposons [23,26]. Here, we performed analyses of the presence and localization of PQSs in the *P. sativum* genome, including its linear nuclear chromosomes and its mitochondrial (mt) and chloroplast (cp) DNA, and we found experimental evidence that sequences found in these two organelles adopted G4 structures in vitro.

2. Results

2.1. Comparison of PQS Sequences in *P. sativum* Genome

The fully sequenced genome of *P. sativum* in the NCBI database consists of seven chromosomes, mitochondrial DNA (mtDNA), and chloroplast DNA (cpDNA). The length of *P. sativum* chromosomes varies between 372 Mbp for chromosome I and 580 Mbp for chromosome V. *P. sativum* mtDNA is 363,843 bp long and cpDNA is 122,035 bp long. G4Hunter analyses with standard values for G4Hunter (i.e., a window size of 25 nucleotides and threshold score of 1.2), showed over 1.3 million PQSs in *P. sativum* genome (Table 1).

In total, we found 1,355,394 PQSs, with no obvious strand bias (679,713 in one strand and 675,681 in the complementary strand). Detailed results for each sequence are presented in the Supplementary Materials (Table S1). As expected, the most abundant PQSs had a moderate G4Hunter score (G4HS) in the 1.2–1.4 category (70.1% of all PQSs), followed by sequences in the 1.4–1.6 (19.2% of all PQSs) and 1.6–1.8 (5.7% of all PQSs) intervals. Sequences with a high G4HS (1.8–2.0 interval: 28,513 PQSs; 2.0–more: 28,801 PQSs) were the least frequent. As expected, the number of PQSs tended to decrease with the G4Hunter threshold. The frequency of the PQSs with the G4Hunter score in the 1.2–1.6 interval was higher in mtDNA than in nuclear DNA and cpDNA. We compared the distribution of G4HSs in *P. sativum* with the relative frequencies of PQSs in various organisms (Figure 1). While the genome of *P. sativum* contained more PQSs with a G4HS above 1.8 and 2.0 compared to prokaryotic genomes of *Escherichia coli* (bacteria) and *Haloferax volcanii* (Archaea), the number of PQSs in these categories was higher in animals such as *C. elegans* and *H. sapiens*.

Table 1. Total number and frequencies of PQSs found in *P. sativum* genome grouped according to G4Hunter score (1.2–1.4 means any sequence with a score between 1.2 and 1.399; 1.4–1.6 between 1.4 and 1.599, etc.).

G4Hunter Threshold	Number of PQSs	PQS Frequency (PQS/kbp)
Genomic DNA		
1.2–1.4	960,462	0.30
1.4–1.6	260,428	0.081
1.6–1.8	76,552	0.024
1.8–2.0	28,513	0.0088
2.0–more	28,801	0.0089
mtDNA		
1.2–1.4	377	1.04
1.4–1.6	117	0.32
1.6–1.8	47	0.13
1.8–2.0	16	0.044
2.0–more	16	0.044
cpDNA		
1.2–1.4	40	0.33
1.4–1.6	15	0.12
1.6–1.8	8	0.066
1.8–2.0	1	0.0082
2.0–more	1	0.0082

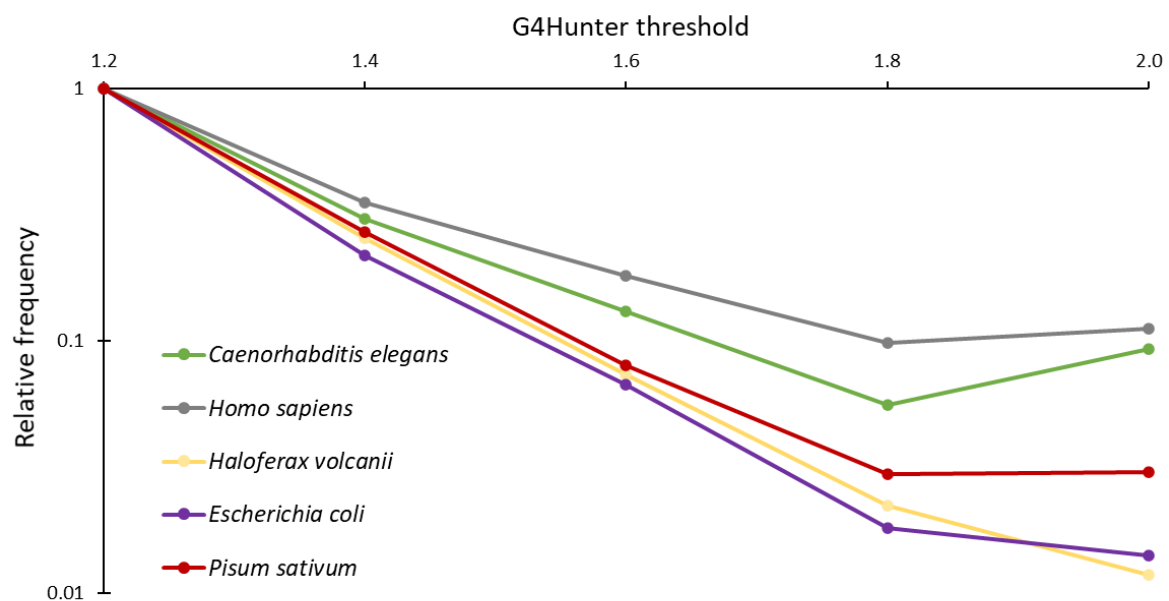


Figure 1. Comparison of G4Hunter score distribution across the different phylogenetic groups. Note the stronger counterselection against high-stability G4s in prokaryotes. *P. sativum*, with an initial slope closer to prokaryotes than to the two other eukaryotes studied here, exhibited an increase in PQS frequency with the highest analyzed G4Hunter score.

We then performed analyses using GC content as an additional parameter to evaluate the influence of GC content on PQS density. The average GC content in the nuclear DNA was 30.02%, with a minimum of 29.68% for chromosome II and a maximum of 31.07% for chromosome I. The frequency of PQSs per 1000 GC for genomic DNA was 1.395. The highest GC content (45.07%) and the highest number of PQSs per 1000 GC (3.494) were found in mtDNA. Chloroplast DNA contained 1.531 PQSs per 1000 GC pairs, with a GC content of 34.78%. However, the frequency of PQSs in chromosomal DNA was very similar for all

chromosomes, and varied between 0.411 PQS per kbp for chromosome II and 0.432 PQS per kbp for chromosome I, with an average of 0.419 PQS/kbp. The frequency of PQSs in cpDNA was 0.533 PQS/kbp (i.e., slightly higher than in nuclear DNA), while the highest PQS frequency was found in mtDNA (1.575 PQSs per 1000 nucleotides, corresponding to 573 PQSs in a 364 kbp genome). In other words, the density in PQSs was nearly four times higher in mtDNA than in nuclear DNA. The total PQS counts and the percentage of GC and PQS frequency characteristics for each sequence are summarized in Table 2.

Table 2. The overall number of PQSs found with a G4Hunter score of 1.2 or above; their frequencies per kbp; GC content; length of all PQSs (all base pairs with potential to form G4) divided by the total number of bp in the DNA (PQSs); and the number of PQSs per thousand GC for each chromosome, mtDNA, and cpDNA.

DNA Sequence	Length (Mb)	Number of PQS	PQS Frequency (/kbp)	GC Content (%)	PQSs (%)	PQSs/GC%
Chr I	372.17	160,922	0.432	31.07	1.31	1.392
Chr II	427.60	175,744	0.411	29.68	1.24	1.385
Chr III	437.56	181,878	0.416	29.72	1.26	1.399
Chr IV	446.35	184,737	0.414	29.90	1.25	1.384
Chr V	579.27	244,737	0.422	30.13	1.28	1.402
Chr VI	480.42	200,963	0.418	29.81	1.27	1.403
Chr VII	491.38	205,775	0.419	29.87	1.27	1.402
Total nuclear	3234.74	1,354,756	0.419	30.02	1.27	1.395
mtDNA	0.36	573	1.575	45.07	4.81	3.494
cpDNA	0.12	65	0.533	34.78	1.65	1.531

2.2. Experimental Demonstration of G4 Formation for Pisum mtDNA and cpDNA Sequences

Among the 573 mtDNA and 65 cpDNA PQSs, we chose 12 candidate sequences (six mtDNA and six cpDNA) spanning G4H scores between 1.25 and 2.0 (Table 3) and representative of motifs found in these organelles. We used a combination of biophysical methods to confirm G4 formation in vitro, as illustrated in Figure 2. As inferred from isothermal difference spectra (IDS) (Figure 2A,B), circular dichroism (CD) spectra (Figure 2C,D), and FRET-MC (Figure 2E,F) for most (11/12) motifs clearly formed G4s at room temperature, while some ambiguity remained for 28ps2. Of note, the majority of spectra (8 out of 12) suggested a parallel fold. This bias was the result of relatively high G4Hunter scores (average G4H = 1.61) and the fact that we systematically introduced non-G nucleotides at both extremities, as flanking nucleotides favor a parallel topology [27].

Table 3. Twelve sequences were analyzed using three different biophysical methods (IDS: isothermal difference spectra; CD: circular dichroism; FRET-MC, a competition fluorescence melting assay). G4Hunter score is indicated in the column labeled “G4H”. Concl. column indicates the conclusion reached based on these three methods. “+” stands for positive, meaning that the method indicated the sequence was forming a G4.

Name	Sequence	G4H	IDS	CD	FRET-MC	Concl.
Mitochondrial sequences:						
40ps1	TGGGCGTCTGGGGTTGGTTTAAAGGAAAAATCGGGGTCGGA	1.25	+	+	+	G4
28ps2	AGGGATCAAGAAACGGATAGGGAGGGGA	1.32	?	+	-	G4?
37ps3	AGGGAGGACCGGGGGCCAGAGCAAGTTGGGTTGGGGT	1.41	+	+	+	G4
44ps4	TGGGGCGAGGGTCTTTCATTAAAGGGGGGAAAAGAGGGGTGGGT	1.66	+	+	+	G4
28ps5	CGGGGGCGGGTCTGAGCAGGATGGGGGA	1.68	+	+	+	G4
31ps6	AGGAAGCGGGGGGAGGAACACAGGGGAAGGA	1.61	+	+	+	G4

Table 3. Cont.

Name	Sequence	G4H	IDS	CD	FRET-MC	Concl.
Chloroplast sequences:						
28ps16	TGGAAGGGGTCAATAAGGGGTGGGGGA	1.96	+	+	+	G4
32ps17	CGGGGGTAGATTGGGGCGTGGACATAAGGGT	1.62	+	+	+	G4
25ps18	TGGGATCCGGGGCGGTCCAGGGGGGA	1.48	+	?	+	G4
24ps23	AGGGGTGGGGACAGAGGTTTTGGT	1.67	+	+	+	G4
21ps26	TGGGGTGGTGAAGGGAGGGC	2.00	+	+	+	G4
24ps27	CGGGGTGGAGACGATGGGGTCGGT	1.62	+	?	+	G4

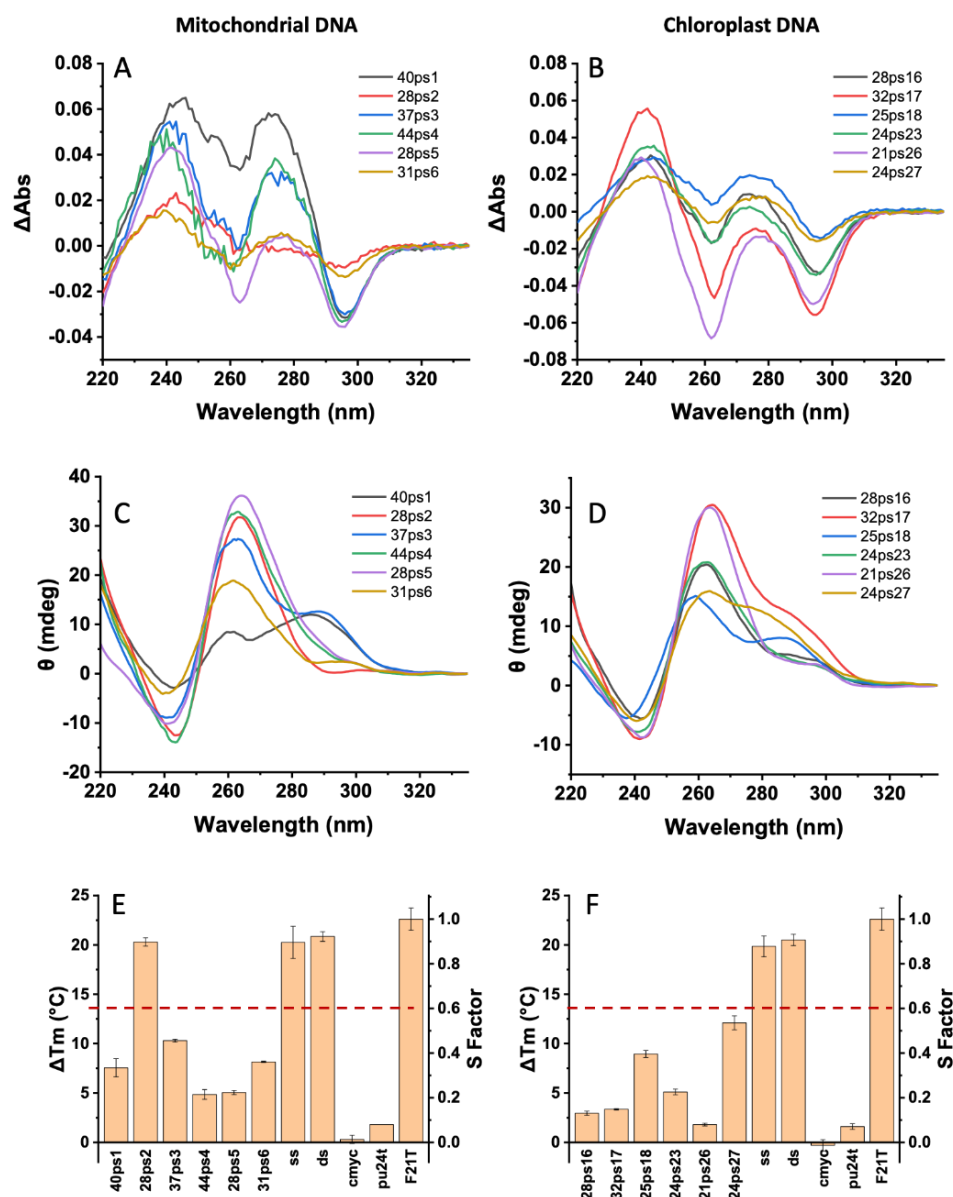


Figure 2. Experimental evidence for G4 formation. (A,B) Isothermal difference spectra (IDS); (C,D) circular dichroism spectra; (E,F) FRET-MC results for the mitochondrial (left) and chloroplast (right) sequences. In panels E and F, ss and ds correspond to single- and double-stranded negative controls, while cmyc and pu24t are G4-forming positive controls. F21T corresponds to the ΔTm observed in the absence of any competitor ($S = 1$). The red dotted line corresponds to the threshold under which a sequence was considered to form a quadruplex [15,28].

2.3. Localization of PQSs in *P. sativum* Genome

To analyze the localization of PQSs in *P. sativum*, we downloaded annotations from the NCBI genome database and overlaid the PQS presence with described *features* and repeats identified de novo in RepeatExplorer2. In addition to the direct presence within the *features*, we also analyzed the presence of PQSs 100 bp before and after the *feature* annotations (Figure 3; Table S2 in the Supplementary Materials). An analysis of PQSs in annotated features and repeats showed that the distribution of PQSs throughout the genome was not uniform. The highest frequency of PQSs per kbp in genomic DNA was found within coding regions (CDS (0.785)), and around *repeat regions* ((0.580)—retrotransposons and transposons) within the *mRNA* (0.534). A notable enrichment in PQSs was also found within 5'UTR, while few PQSs were present in 3'UTR. The lowest PQS frequency was found before or after *ncRNA* (0.23 and 0.16) and *within* 3'UTR (0.287). The density around 3'UTR (0.315) was also lower than average. The annotations for genomic DNA are shown in Figure 3A. The telomeric motif of *P. sativum* has been known for a long time [29]; however, the telomeres were not annotated in the current assembly. The telomeric repeats of *P. sativum* were composed of TTTAGGG repeats, which had a G4Hunter score of 1.29. G4 formation with this motif has been previously demonstrated [30,31], and the stability of the corresponding G4 was relatively high ($T_m = 64\text{ }^\circ\text{C}$ in 100mM KCl; nearly as high as the human hexanucleotide GGGTTA motif). There were ≈ 142 TTTAGGG motifs per kbp of telomeric DNA, which would allow the formation of up to 35–36 G4s per kbp, but the formation of multiple juxtaposed G4s has not been experimentally investigated for this plant motif.

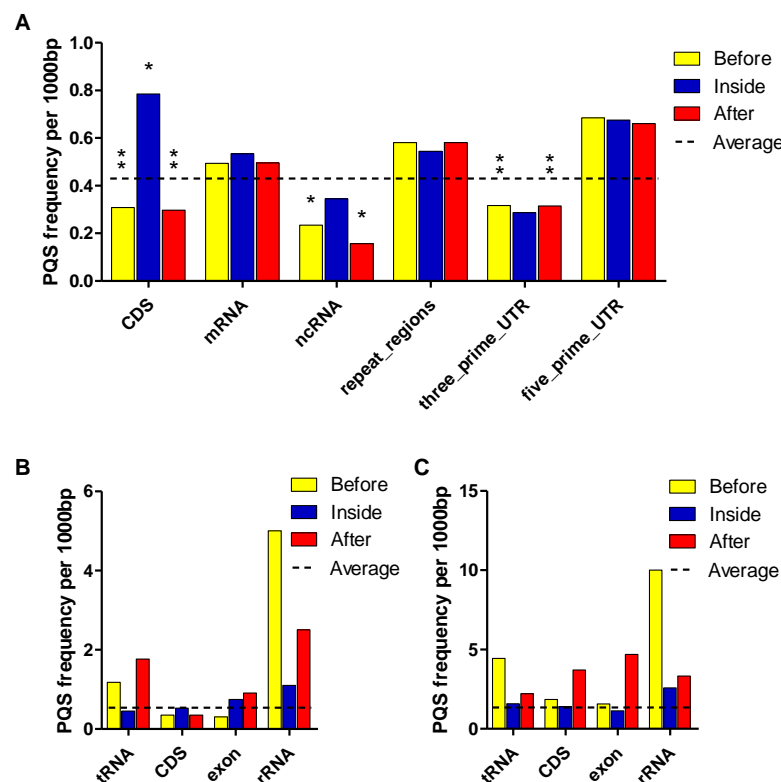


Figure 3. Differences in PQS frequency according to DNA locus. The chart shows PQS frequencies normalized per 1000 bp annotated locations from the NCBI database. We analyzed the frequencies of all PQSs within (inside), before (100 bp), and after (100 bp) annotated locations in (A) genomic DNA, (B) mtDNA, and (C) cpDNA. Dashed lines denote the average PQS frequency in corresponding DNA. Statistical significance of annotated locations in genomic DNA was related to the average chromosomal PQS frequencies according to a Kruskal–Wallis test, followed by Dunn’s pairwise comparison with Bonferroni correction of the *p*-value. Asterisks denote statistical significance: * *p*-value < 0.05; ** *p*-value < 0.01.

In mtDNA, the highest PQS frequencies per kbp were found within 100 bp after exon (4.688), before tRNA (4.444), and after CDS (3.703), followed by region 100 bp after (3.333) and inside rRNA (2.575). The most notable enrichment of PQSs was found in the regions before rRNA, where the PQS frequency per 1000 bp reached 10. The lowest PQS frequency was found within the exon region (1.130). Differences in PQS frequency according to annotated features in mtDNA are shown in Figure 3B.

In cpDNA, a high PQS frequency within features was also observed 100 bp before (5.0) and after (2.5) rRNA, similarly to mtDNA. The frequency inside rRNA was almost 5-times lower than the frequency before this feature (see Figure 3C).

2.4. PQSs in Transposable Elements

Using the G4Hunter algorithm, we analyzed *P. sativum* repeat regions to determine the frequency and distribution of PQSs in transposable elements (TEs). In the case of the *P. sativum* genome, TEs represented over 80%, with a significant contribution by Ogre elements, which is a group of LTR retrotransposons (Class I) [32]. LTR retrotransposons of the superfamilies Ty3-gypsy and Ty1-copia were the dominant group, with over 91% and over 8.5% of the LTR sequence coverage, respectively. The transposons (Class II) represented a smaller part of the genome. Over 99% of all Class II transposons were terminal-inverted repeat (TIR) transposons, and less than 1% were helitrons. Satellite and ribosomal DNA (rDNA) formed a small fraction of all annotated TEs. Short tandem repeats annotated as Pararetrovirus were the result of the viral sequence integration [33]. When only total PQSs were considered, the largest number of PQSs in annotated transposons was found within Ty3/gypsy, Ty1/copia, and 100 bp before/after Ty3/gypsy. To evaluate the localization of PQSs within TEs, we overlapped PQSs with annotated locations and analyzed the frequencies of all PQSs within, 100 bp before, and 100 bp after annotated TEs (Figure 4). The only TEs with a higher frequency of PQSs inside than before and after were unclassified transposons and Ty1-copia.

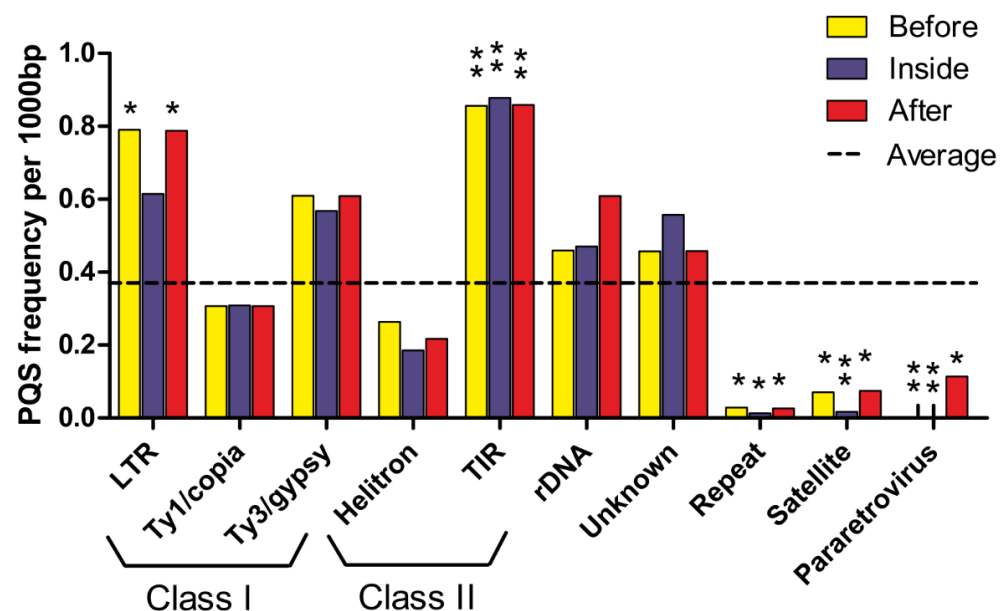


Figure 4. Differences in PQS frequency by repeat region. The chart shows PQS frequencies normalized per 1000 bp of annotated transposons. We analyzed the frequencies of all PQSs within (inside), before (100 bp), and after (100 bp) annotated transposons in genomic DNA. The dashed line denotes the average PQS frequency in transposons. Statistical significance is shown as in Figure 3. $p < 0.05$, * $p < 0.01$, ** $p < 0.001$.

The highest PQS frequency per kbp was observed within TIR transposons. High PQS frequencies were found before and after unclassified LTRs (almost 1.5-times more frequent than in gene regions). Repeats and satellites had the lowest PQS frequencies per kbp, and

their PQS frequencies compared to the gene region were 5 times less frequent than in gene regions. No PQSs were found at 100 bp before and inside Pararetrovirus. However, this was not the same when considering PQS frequency per kbp after Pararetrovirus. The data are available in Table S4 in the Supplementary Materials.

3. Discussion

The recent improvements in the sequencing methods of and computational approaches to full-genome analyses allowed effective searches for PQSs. The G4Hunter algorithm was successfully used to select PQSs with a high probability of G4 structure formation and minimum positive or false negative results in various genomes from viruses, bacteria, and eukaryotes, including the human genome.

However, the number of plant genomes analyzed for G4 propensity is still limited, and is mostly performed using older pattern-based algorithms. There are several reasons for this: the plant genomes are usually huge, and compared to an animal, there are not so many fully assembled plant genomes; moreover, the number of repetitive sequences in some plant genomes is enormous, and these repetitive sequences are challenging for the correct assembly in the genomes. Genomic DNA from *P. sativum* contained various repetitive sequences involving transposable elements (TE). Previous studies showed that the TE fraction represented a significant portion of plant genomes, and could vary from 15 to 30% in *Arabidopsis thaliana* (thale cress) and *Brachypodium distachyon* (purple false brome), and from 70 to 80% in species such as *Zea mays* ssp. *mays* (maize) and *Hordeum vulgare* (barley) [34]. In the case of the *P. sativum* genome, TEs represent more than 80% [32]. Therefore, we took advantage of the contemporary sequenced genome of *P. sativum* and performed G4Hunter analyses to determine the presence and localization of PQSs within classic features, as well as TEs. PQSs have been identified in various plant genomes, including *A. thaliana*, *Oryza sativa* subsp. *Japonica* (rice), *Populus trichocarpa* (black cottonwood), and *Vitis vinifera* (common grape) [35]. Previous pattern-based PQS analyses (Quadparser G3L1-7; corresponding to a motif involving four runs of at least three guanines separated by loops of one to seven nucleotides) demonstrated that *Arabidopsis* had only 9 G4 motifs/Mbp, while rice had 92 G4 motifs/Mbp, a 10-fold difference, and the monocot plant sample (barley, maize, and rice) had a higher PQS frequency compared to dicots (soybean, common grapevine, and *Arabidopsis thaliana*) [21,36,37]. It is hard to compare various algorithms for PQS prediction; however, considering only PQSs with a G4Hunter score above 1.4, which represents a very stable G4 as evaluated in vitro, the frequency of PQSs in *P. sativum* genome seemed higher than within previously reported dicot plants.

However, the more interesting aspect was the huge difference in PQSs between nuclear and organelle DNA, especially mtDNA. While the frequency of PQSs in all chromosomes was similar (around 0.41 per kbp), mtDNA had more than five times as many PQSs, suggesting a different regulation for *P. sativum* linear nuclear and circular mtDNA. PQSs also had different localizations in mtDNA compared to PQS localization in nuclear DNA; therefore, we suggest that G4-formation and regulatory pathways differ in circular and linear DNAs. Interestingly the comparison of mtDNA PQS frequencies among various species showed an increased PQS frequency for vertebrates as well as for land plants, contrary to a lower PQS frequency in the mtDNA of protists and fungi [38]. In animals, it has been shown that G4s play a direct role in mitochondrial genome replication, transcription processivity, and respiratory function [39]. The significantly higher frequency of PQSs in *P. sativum* mtDNA compared to those in nuclear DNA suggested that this observation may also be valid for plant mitochondria. We analyzed a dozen sequences in vitro that were extracted from *P. sativum* mitochondrial (mt) and chloroplast (cp) DNA, and provided experimental evidence that the motifs found in these two organelles were prone to G4 formation in vitro. This study constitutes, to the best of our knowledge, the first experimental evidence that chloroplast sequences may form G4s.

Generally, it is accepted that very stable G4s tend to be strongly counter-selected, and low-scoring PQSs (with a G4HS between 1.2 and 1.6) tend to constitute the vast majority

of G4-prone motifs. The main reason is probably the high stability of G4s formed by PQSs with a G4HS > 1.8, which therefore constitutes a physical barrier for most biological processes such as replication or transcription [40]. Interestingly, there was no significant drop in PQS density in the *P. sativum* genome for high G4H scores. This was in contrast with what has been observed in most other species, as previous analyses revealed that most of the PQSs found in Platyhelminthes [41], Archaea [15], and bacteria [14] have a relatively low G4Hunter score, and the number of PQSs in these organisms decreases sharply above a score of 1.6. Strikingly, nuclear DNA and mt- and cpDNA differ not only in PQS frequencies, but also in the localization of these PQSs. For circular cp- and mtDNA, there was a strong abundance around RNA genes, while in genomic DNA, there was a significant difference in PQS frequency for 3'UTR, where PQS were more than twice as less frequent inside compared to 5'UTR. The 5'UTR serves as the binding point for the ribosome, which allows the ribosome to bind and initiate translation [42]. The higher density of PQSs in this region suggested important regulatory roles of G4 motifs in the process of translation. Many G4-binding proteins in animals and humans are known [43]. Recently, it has been shown that proteins in barley seedlings can bind to PQSs and form DNA–protein complexes [44], so we can expect that G4-binding proteins also will be present in plant genomes [45].

In conclusion, we analyzed the presence of PQSs in cpDNA and TE for the first time. The nonrandom localization of PQSs in the genome of *P. sativum* suggested their regulatory function and the importance of LTR and TIR transposons. This supported the hypothesis that TEs may serve as vehicles for the genomic spread of G4s [46]. In addition, the higher density of PQSs in mtDNA and cpDNA compared to regular chromosomes suggested specific roles for quadruplexes in organelles.

4. Materials and Methods

4.1. Process of Analysis

The complete DNA sequences of the *P. sativum* genome, including nuclear, mt, and cp genomes were downloaded (20 June 2021) in FASTA format from the National Center for Biotechnology Information (NCBI) [47]. NCBI IDs are listed in Table S1 in the Supplementary Materials. For putative PQS prediction, the new and strengthened computational core of our DNA analyzer software written in Java programming language was used [48]. For our analyses, we used an actualized G4Hunter algorithm implementation [49] with default parameters for G4Hunter—a window size of 25 and a G4Hunter score (G4HS) above 1.2 (the chosen value of 25 nucleotides corresponded to the size of a typical intramolecular G4). The default values for G4Hunter have been previously discussed and validated [50]. G4HSs were then grouped in five intervals: 1.2 up to 1.4, 1.4 up to 1.6, 1.6 up to 1.8, 1.8 up to 2.0, and 2.0 and more. Data were merged in a single Excel file (accessible in Table S1 in the Supplementary Materials) for further analyses and statistical evaluations.

4.2. Analysis of Repetitive DNA from Unassembled Reads Using RepeatExplorer2 and TAREAN

Only the conserved coding domains of the repeats were annotated in the available genome assembly [23]. Therefore, we performed an independent de novo identification of repeats and annotated genomic loci corresponding to them, including their specific regions such as LTRs, spacer sequences, etc. We used publicly available low-pass whole-genome sequencing data in FASTQ format from the Sequence Read Archive of the NCBI (Run ERR063464) [51]. We performed standard preprocessing, a quality check, and interlacing of paired-end reads, and ran RepeatExplorer2 and TAREAN analyses with 2,913,990 reads of a uniform length of 100 nt [52]. The results were manually checked, and the sequences of selected repeats—mobile elements (LTR, Ty1/copia, Ty3/gypsy, TIR, helitron, pararetrovirus), satellites, rDNA, and unclassified repeats—were used in BLAST against the pea genome for the purpose of repeat loci annotation according to the feature table (see below) completion.

4.3. Sequence Matching and Transposon Annotation (BLAST)

The BLAST database was constructed from the pea genomic sequence (accessible in Table S1 in the Supplementary Materials), and the sequences from our RepeatExplorer2 analysis (see Section 2.2) were used as a query in blastn with parameters as follows: -outfmt 6-max_target_seqs 10000000-num_threads 4-evalue 0.1. The blast match positions were then used for the feature table completion.

4.4. Analysis of PQSs around Annotated NCBI Features and Repeats from Our RepeatExplorer2 Analysis

The feature table containing functional annotations of the *P. sativum* genome was downloaded from the NCBI database. Features describe the functions and locations of sequences within the genome of an organism [53]. We performed an analysis of PQS occurrence inside uploaded features, as well as 100 bp before and after each feature. Features were grouped by their name stated in the feature table file (gene, rRNA, tRNA, ncRNA, and repeat region). Further processing was performed in Microsoft Excel, and the resulting data are available in Table S2 in the Supplementary Materials.

4.5. Statistical Analysis

Outliers were detected using the function `chisq.out.test` from the outliers package in R version 4.0.5 [54]. Normal distribution of PQS frequencies in annotated locations was determined using the Shapiro–Wilk test, and statistical significance was evaluated using the nonparametric Kruskal–Wallis test. Multiple pairwise comparisons were assessed using a post hoc Dunn’s test with Bonferroni correction of the significance level.

4.6. Experimental Demonstration of G4 Formation

DNA sequence matching motifs found in *Pisum* chloroplast and mitochondrial DNA were synthesized by Eurogentec (Seraing, Belgium) and used without further purification. Concentrations were determined using the extinction coefficients provided by the manufacturer. Isothermal difference spectra (IDS) and circular dichroism (CD) spectra were recorded as previously described [28]. FRET-MC provided a convenient independent method to detect G4 formation; detailed experimental protocols can be found in [15,28]. Briefly, in this test, G4-forming competitors led to a marked decrease in the ligand-induced stabilization effect (ΔT_m), while nonspecific competitors (e.g., single- or double-stranded sequences) had little effect.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/ijms23158482/s1>.

Author Contributions: Conceptualization, V.B. and M.B.; software, formal analysis, resources, M.D., N.B. and V.P.; visualization, M.D. and Y.L.; validation, N.B.; investigation, M.D., N.B., V.P., Y.L. and J.W.; writing—review and editing, J.-L.M., M.D., A.V., V.B. and V.P.; validation, N.B.; supervision, V.B. and J.-L.M. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the SYMBIT project (Reg. No. CZ.02.1.01/0.0/0.0/15_003/0000477; financed by the ERDF), the University of Ostrava (SGS11/PřF/2022), and by the Czech Science Foundation (No. 22-21903S, 21-18532S).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data are available in the paper and supplementary materials.

Acknowledgments: We thank A. Cucchiari, L. Guittat (LOB) and D. Verga (Institut Curie) for helpful discussions.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Trněný, O.; Brus, J.; Hradilová, I.; Rathore, A.; Das, R.R.; Kopecký, P.; Coyne, C.J.; Reeves, P.; Richards, C.; Smýkal, P. Molecular Evidence for Two Domestication Events in the Pea Crop. *Genes* **2018**, *9*, 535. [[CrossRef](#)] [[PubMed](#)]
2. Powers, S.E.; Thavarajah, D. Checking Agriculture's Pulse: Field Pea (*Pisum Sativum* L.), Sustainability, and Phosphorus Use Efficiency. *Front. Plant Sci.* **2019**, *10*, 1489. [[CrossRef](#)] [[PubMed](#)]
3. Gu, B.; Chen, Y.; Xie, F.; Murray, J.D.; Miller, A.J. Inorganic Nitrogen Transport and Assimilation in Pea (*Pisum Sativum*). *Genes* **2022**, *13*, 158. [[CrossRef](#)] [[PubMed](#)]
4. Labeeb, M.; Badr, A.; Haroun, S.A.; Mattar, M.Z.; El-Kholy, A.S. Ultrastructural and Molecular Implications of Ecofriendly Made Silver Nanoparticles Treatments in Pea (*Pisum Sativum* L.). *J. Genet. Eng. Biotechnol.* **2022**, *20*, 5. [[CrossRef](#)]
5. Mendel, G.J. Versuche Über Pflanzenhybriden. *Verh. Nat. Ver. Brünn Abh.* **1865**, *4*, 3–47.
6. Bateson, W. *Mendel's Principles of Heredity*; Cambridge University Press: Cambridge, UK, 1902; ISBN 978-0-511-69446-2.
7. Bartas, M.; Brázda, V.; Karlický, V.; Červeň, J.; Pečinka, P. Bioinformatics Analyses and in Vitro Evidence for Five and Six Stacked G-Quadruplex Forming Sequences. *Biochimie* **2018**, *150*, 70–75. [[CrossRef](#)]
8. Cho, H.; Cho, H.S.; Nam, H.; Jo, H.; Yoon, J.; Park, C.; Dang, T.V.T.; Kim, E.; Jeong, J.; Park, S.; et al. Translational Control of Phloem Development by RNA G-Quadruplex–JULGI Determines Plant Sink Strength. *Nat. Plants* **2018**, *4*, 376–390. [[CrossRef](#)]
9. Kim, N. The Interplay between G-Quadruplex and Transcription. *Curr. Med. Chem.* **2019**, *26*, 2898–2917. [[CrossRef](#)]
10. Robinson, J.; Raguseo, F.; Nuccio, S.P.; Liano, D.; Di Antonio, M. DNA G-Quadruplex Structures: More than Simple Roadblocks to Transcription? *Nucleic Acids Res.* **2021**, *49*, 8419–8431. [[CrossRef](#)]
11. Feng, Y.; Tao, S.; Zhang, P.; Sperti, F.R.; Liu, G.; Cheng, X.; Zhang, T.; Yu, H.; Wang, X.-E.; Chen, C.; et al. Epigenomic Features of DNA G-Quadruplexes and Their Roles in Regulating Rice Gene Transcription. *Plant Physiol.* **2022**, *188*, 1632–1648. [[CrossRef](#)]
12. Bohálová, N.; Cantara, A.; Bartas, M.; Kaura, P.; Šťastný, J.; Pečinka, P.; Fojta, M.; Mergny, J.-L.; Brázda, V. Analyses of Viral Genomes for G-Quadruplex Forming Sequences Reveal Their Correlation with the Type of Infection. *Biochimie* **2021**, *186*, 13–27. [[CrossRef](#)]
13. Lavezzo, E.; Berselli, M.; Frasson, I.; Perrone, R.; Palù, G.; Brazzale, A.R.; Richter, S.N.; Toppo, S. G-Quadruplex Forming Sequences in the Genome of All Known Human Viruses: A Comprehensive Guide. *PLoS Comput. Biol.* **2018**, *14*, e1006675. [[CrossRef](#)]
14. Bartas, M.; Čutová, M.; Brázda, V.; Kaura, P.; Šťastný, J.; Kolomazník, J.; Coufal, J.; Goswami, P.; Červeň, J.; Pečinka, P. The Presence and Localization of G-Quadruplex Forming Sequences in the Domain of Bacteria. *Molecules* **2019**, *24*, 1711. [[CrossRef](#)]
15. Brázda, V.; Luo, Y.; Bartas, M.; Kaura, P.; Porubiaková, O.; Šťastný, J.; Pečinka, P.; Verga, D.; Da Cunha, V.; Takahashi, T.S. G-Quadruplexes in the Archaea Domain. *Biomolecules* **2020**, *10*, 1349. [[CrossRef](#)]
16. Čutová, M.; Manta, J.; Porubiaková, O.; Kaura, P.; Šťastný, J.; Jagelská, E.B.; Goswami, P.; Bartas, M.; Brázda, V. Divergent Distributions of Inverted Repeats and G-Quadruplex Forming Sequences in *Saccharomyces Cerevisiae*. *Genomics* **2020**, *112*, 1897–1901. [[CrossRef](#)]
17. Warner, E.F.; Bohálová, N.; Brázda, V.; Waller, Z.A.E.; Bidula, S. Analysis of Putative Quadruplex-Forming Sequences in Fungal Genomes: Novel Antifungal Targets? *Microb. Genom.* **2021**, *7*, 000570. [[CrossRef](#)]
18. Hänsel-Hertsch, R.; Di Antonio, M.; Balasubramanian, S. DNA G-Quadruplexes in the Human Genome: Detection, Functions and Therapeutic Potential. *Nat. Rev. Mol. Cell Biol.* **2017**, *18*, 279–284. [[CrossRef](#)]
19. Garg, R.; Aggarwal, J.; Thakkar, B. Genome-Wide Discovery of G-Quadruplex Forming Sequences and Their Functional Relevance in Plants. *Sci. Rep.* **2016**, *6*, 28211. [[CrossRef](#)]
20. Yang, X.; Cheema, J.; Zhang, Y.; Deng, H.; Duncan, S.; Umar, M.I.; Zhao, J.; Liu, Q.; Cao, X.; Kwok, C.K. RNA G-Quadruplex Structures Exist and Function in Vivo in Plants. *Genome Biol.* **2020**, *21*, 226. [[CrossRef](#)]
21. Griffin, B.D.; Bass, H.W. Plant G-Quadruplex (G4) Motifs in DNA and RNA. Abundant, Intriguing Sequences of Unknown Function. *Plant Sci.* **2018**, *269*, 143–147. [[CrossRef](#)]
22. Volná, A.; Bartas, M.; Karlický, V.; Nezval, J.; Kundrátová, K.; Pečinka, P.; Špunda, V.; Červeň, J. G-Quadruplex in Gene Encoding Large Subunit of Plant RNA Polymerase II: A Billion-Year-Old Story. *Int. J. Mol. Sci.* **2021**, *22*, 7381. [[CrossRef](#)] [[PubMed](#)]
23. Kreplak, J.; Madoui, M.-A.; Čápal, P.; Novák, P.; Labadie, K.; Aubert, G.; Bayer, P.E.; Gali, K.K.; Syme, R.A.; Main, D.; et al. A Reference Genome for Pea Provides Insight into Legume Genome Evolution. *Nat. Genet.* **2019**, *51*, 1411–1422. [[CrossRef](#)] [[PubMed](#)]
24. Ellis, T.H.N.; Poyser, S.J. An Integrated and Comparative View of Pea Genetic and Cytogenetic Maps. *New Phytol.* **2002**, *153*, 17–25. [[CrossRef](#)]
25. Macas, J.; Novák, P.; Pellicer, J.; Čížková, J.; Koblížková, A.; Neumann, P.; Fuková, I.; Doležel, J.; Kelly, L.J.; Leitch, I.J. In Depth Characterization of Repetitive DNA in 23 Plant Genomes Reveals Sources of Genome Size Variation in the Legume Tribe Fabaeae. *PLoS ONE* **2015**, *10*, e0143424. [[CrossRef](#)]
26. Li, S.-F.; Su, T.; Cheng, G.-Q.; Wang, B.-X.; Li, X.; Deng, C.-L.; Gao, W.-J. Chromosome Evolution in Connection with Repetitive Sequences and Epigenetics in Plants. *Genes* **2017**, *8*, 290. [[CrossRef](#)]
27. Chen, J.; Cheng, M.; Salgado, G.F.; Stadlbauer, P.; Zhang, X.; Amrane, S.; Guédin, A.; He, F.; Šponer, J.; Ju, H.; et al. The Beginning and the End: Flanking Nucleotides Induce a Parallel G-Quadruplex Topology. *Nucleic Acids Res.* **2021**, *49*, 9548–9559. [[CrossRef](#)]
28. Luo, Y.; Granzhan, A.; Verga, D.; Mergny, J.-L. FRET-MC: A Fluorescence Melting Competition Assay for Studying G4 Structures in Vitro. *Biopolymers* **2021**, *112*, e23415. [[CrossRef](#)]

29. Cesare, A.J.; Quinney, N.; Willcox, S.; Subramanian, D.; Griffith, J.D. Telomere Looping in *P. sativum* (Common Garden Pea). *Plant J.* **2003**, *36*, 271–279. [[CrossRef](#)]
30. Tran, P.L.T.; Mergny, J.-L.; Alberti, P. Stability of Telomeric G-Quadruplexes. *Nucleic Acids Res.* **2011**, *39*, 3282–3294. [[CrossRef](#)]
31. De Cian, A.; Grellier, P.; Mouray, E.; Depoix, D.; Bertrand, H.; Monchaud, D.; Teulade-Fichou, M.-P.; Mergny, J.-L.; Alberti, P. Plasmodium Telomeric Sequences: Structure, Stability and Quadruplex Targeting by Small Compounds. *ChemBioChem* **2008**, *9*, 2730–2739. [[CrossRef](#)]
32. Burstin, J.; Kreplak, J.; Macas, J.; Lichtenzveig, J. *Pisum Sativum* (Pea). *Trends Genet.* **2020**, *36*, 312–313. [[CrossRef](#)]
33. Jakowitsch, J.; Mette, M.F.; van der Winden, J.; Matzke, M.A.; Matzke, A.J.M. Integrated Pararetroviral Sequences Define a Unique Class of Dispersed Repetitive DNA in Plants. *Proc. Natl. Acad. Sci. USA* **1999**, *96*, 13241–13246. [[CrossRef](#)]
34. Bennetzen, J.L.; Wang, H. The Contributions of Transposable Elements to the Structure, Function, and Evolution of Plant Genomes. *Annu. Rev. Plant Biol.* **2014**, *65*, 505–530. [[CrossRef](#)]
35. Takahashi, H.; Nakagawa, A.; Kojima, S.; Takahashi, A.; Cha, B.-Y.; Woo, J.-T.; Nagai, K.; Machida, Y.; Machida, C. Discovery of Novel Rules for G-Quadruplex-Forming Sequences in Plants by Using Bioinformatics Methods. *J. Biosci. Bioeng.* **2012**, *114*, 570–575. [[CrossRef](#)]
36. Yadav, V.; Kim, N.; Tuteja, N.; Yadav, P. G Quadruplex in Plants: A Ubiquitous Regulatory Element and Its Biological Relevance. *Front. Plant Sci.* **2017**, *8*, 1163. [[CrossRef](#)]
37. Wang, Y.; Zhao, M.; Zhang, Q.; Zhu, G.-F.; Li, F.-F.; Du, L.-F. Genomic Distribution and Possible Functional Roles of Putative G-Quadruplex Motifs in Two Subspecies of *Oryza Sativa*. *Comput. Biol. Chem.* **2015**, *56*, 122–130. [[CrossRef](#)]
38. Bohálová, N.; Dobrovolná, M.; Brázda, V.; Bidula, S. Conservation and Over-Representation of G-Quadruplex Sequences in Regulatory Regions of Mitochondrial DNA across Distinct Taxonomic Sub-Groups. *Biochimie* **2022**, *194*, 28–34. [[CrossRef](#)]
39. Falabella, M.; Kolesar, J.E.; Wallace, C.; de Jesus, D.; Sun, L.; Taguchi, Y.V.; Wang, C.; Wang, T.; Xiang, I.M.; Alder, J.K.; et al. G-Quadruplex Dynamics Contribute to Regulation of Mitochondrial Gene Expression. *Sci. Rep.* **2019**, *9*, 5605. [[CrossRef](#)]
40. Castillo Bosch, P.; Segura-Bayona, S.; Koole, W.; van Heteren, J.T.; Dewar, J.M.; Tijsterman, M.; Knipscheer, P. FANCI Promotes DNA Synthesis through G-Quadruplex Structures. *EMBO J.* **2014**, *33*, 2521–2533. [[CrossRef](#)]
41. Cantara, A.; Luo, Y.; Dobrovolná, M.; Bohalova, N.; Fojta, M.; Verga, D.; Guittat, L.; Cucchiari, A.; Savrimoutou, S.; Häberli, C.; et al. G-Quadruplexes in Helminth Parasites. *Nucleic Acids Res.* **2022**, *50*, 2719–2735. [[CrossRef](#)]
42. Lee, D.S.M.; Ghanem, L.R.; Barash, Y. Integrative Analysis Reveals RNA G-Quadruplexes in UTRs Are Selectively Constrained and Enriched for Functional Associations. *Nat. Commun.* **2020**, *11*, 527. [[CrossRef](#)] [[PubMed](#)]
43. Brázda, V.; Hároníková, L.; Liao, J.C.; Fojta, M. DNA and RNA Quadruplex-Binding Proteins. *Int. J. Mol. Sci.* **2014**, *15*, 17493–17517. [[CrossRef](#)] [[PubMed](#)]
44. Sjakste, T.; Leonova, E.; Petrovs, R.; Trapina, I.; Röder, M.S.; Sjakste, N. Tight DNA-Protein Complexes Isolated from Barley Seedlings Are Rich in Potential Guanine Quadruplex Sequences. *PeerJ* **2020**, *8*, e8569. [[CrossRef](#)] [[PubMed](#)]
45. Volná, A.; Bartas, M.; Nezval, J.; Špunda, V.; Pečinka, P.; Červeň, J. Searching for G-Quadruplex-Binding Proteins in Plants: New Insight into Possible G-Quadruplex Regulation. *BioTech* **2021**, *10*, 20. [[CrossRef](#)] [[PubMed](#)]
46. Kejnovsky, E.; Tokan, V.; Lexa, M. Transposable Elements and G-Quadruplexes. *Chromosome Res.* **2015**, *23*, 615–623. [[CrossRef](#)] [[PubMed](#)]
47. Sayers, E.W.; Agarwala, R.; Bolton, E.E.; Brister, J.R.; Canese, K.; Clark, K.; Connor, R.; Fiorini, N.; Funk, K.; Hefferon, T. Database Resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **2019**, *47*, D23. [[CrossRef](#)] [[PubMed](#)]
48. Brázda, V.; Kolomazník, J.; Lýsek, J.; Hároníková, L.; Coufal, J.; Št'astný, J. Palindrome Analyser—A New Web-Based Server for Predicting and Evaluating Inverted Repeats in Nucleotide Sequences. *Biochem. Biophys. Res. Commun.* **2016**, *478*, 1739–1745. [[CrossRef](#)] [[PubMed](#)]
49. Brázda, V.; Kolomazník, J.; Lýsek, J.; Bartas, M.; Fojta, M.; Št'astný, J.; Mergny, J.-L. G4Hunter Web Application: A Web Server for G-Quadruplex Prediction. *Bioinformatics* **2019**, *35*, 3493–3495. [[CrossRef](#)]
50. Bedrat, A.; Lacroix, L.; Mergny, J.-L. Re-Evaluation of G-Quadruplex Propensity with G4Hunter. *Nucleic Acids Res.* **2016**, *44*, 1746–1759. [[CrossRef](#)]
51. Neumann, P.; Navrátilová, A.; Schroeder-Reiter, E.; Koblížková, A.; Steinbauerová, V.; Chocholová, E.; Novák, P.; Wanner, G.; Macas, J. Stretching the Rules: Monocentric Chromosomes with Multiple Centromere Domains. *PLoS Genet.* **2012**, *8*, e1002777. [[CrossRef](#)]
52. Novák, P.; Neumann, P.; Macas, J. Global Analysis of Repetitive DNA from Unassembled Sequence Reads Using RepeatExplorer2. *Nat. Protoc.* **2020**, *15*, 3745–3776. [[CrossRef](#)]
53. The DDBJ/ENA/GenBank Feature Table Definition | INSDC. Available online: <https://www.insdc.org/documents/feature-table#2> (accessed on 21 March 2022).
54. Komsta, L. Processing Data for Outliers. *R News* **2006**, *6*, 10–13.

References

1. G. Banfalvi, *Ribose Selected as Precursor to Life*. DNA Cell Biol, 2020. **39**(2): p. 177-186.
2. K. Duffy, S. Arangundy-Franklin, and P. Holliger, *Modified nucleic acids: replication, evolution, and next-generation therapeutics*. BMC Biol, 2020. **18**(1): p. 112.
3. J.D. Watson and F.H. Crick, *The structure of DNA*. Cold Spring Harb Symp Quant Biol, 1953. **18**: p. 123-31.
4. J.D. Watson and F.H.C. Crick, *Molecular Structure of Nucleic Acids*. Nature, 1953. **171**(4356): p. 737-738.
5. C. Rivetti, M. Guthold, and C. Bustamante, *Scanning force microscopy of DNA deposited onto mica: equilibration versus kinetic trapping studied by statistical polymer chain analysis*. J Mol Biol, 1996. **264**(5): p. 919-32.
6. R. Rohs, X. Jin, S.M. West, R. Joshi, B. Honig, and R.S. Mann, *Origins of specificity in protein-DNA recognition*. Annu Rev Biochem, 2010. **79**: p. 233-69.
7. C.R. Calladine and H. Drew, *Understanding DNA: the molecule and how it works*. 1997: Academic press.
8. S. Neidle, *Oxford handbook of nucleic acid structure*. 1999: Oxford University Press on Demand.
9. A. Arcella and M. Orozco, *Nucleic Acids in the Gas Phase*. 2013.
10. S. Li, W.K. Olson, and X.J. Lu, *Web 3DNA 2.0 for the analysis, visualization, and modeling of 3D nucleic acid structures*. Nucleic Acids Res, 2019. **47**(W1): p. W26-W34.
11. A.H. Wang, G.J. Quigley, F.J. Kolpak, J.L. Crawford, J.H. van Boom, G. van der Marel, and A. Rich, *Molecular structure of a left-handed double helical DNA fragment at atomic resolution*. Nature, 1979. **282**(5740): p. 680-686.
12. V.K. Subramani, S. Ravichandran, V. Bansal, and K.K. Kim, *Chemical-induced formation of BZ-junction with base extrusion*. Biochemical and Biophysical Research Communications, 2019. **508**(4): p. 1215-1220.
13. P. Khuu, M. Sandor, J. DeYoung, and P.S. Ho, *Phylogenomic analysis of the emergence of GC-rich transcription elements*. Proceedings of the National Academy of Sciences, 2007. **104**(42): p. 16528.
14. S. Ravichandran, V.K. Subramani, and K.K. Kim, *Z-DNA in the genome: from structure to disease*. Biophys Rev, 2019. **11**(3): p. 383-387.
15. J.R. Buzzo, A. Devaraj, E.S. Gloag, J.A. Jurcisek, F. Robledo-Avila, T. Kesler, K. Wilbanks, L. Mashburn-Warren, S. Balu, J. Wickham, L.A. Novotny, P. Stoodley, L.O. Bakaletz, and S.D. Goodman, *Z-form extracellular DNA is a structural component of the bacterial biofilm matrix*. Cell, 2021. **184**(23): p. 5740-5758 e17.
16. K. Minton, *ADAR1 inhibits ZBP1 activation by endogenous Z-RNA*. Nature Reviews Genetics, 2022. **23**(10): p. 581-581.
17. S. Arnott, R. Chandrasekaran, D.W.L. Hukins, P.J.C. Smith, and L. Watts, *Structural details of a double-helix observed for DNAs containing alternating purine and pyrimidine sequences*. Journal of Molecular Biology, 1974. **88**(2): p. 523-533.

18. H. Urata, K. Shinohara, E. Ogura, Y. Ueda, and M. Akagi, *Mirror-image DNA*. Journal of the American Chemical Society, 1991. **113**(21): p. 8174-8175.
19. N.C. Hauser, R. Martinez, A. Jacob, S. Rupp, J.D. Hoheisel, and S. Matysiak, *Utilising the left-helical conformation of L-DNA for analysing different marker types on a single universal microarray platform*. Nucleic Acids Research, 2006. **34**(18): p. 5101-5111.
20. A. Rich, D.R. Davies, F.H.C. Crick, and J.D. Watson, *The molecular structure of polyadenylic acid*. Journal of Molecular Biology, 1961. **3**(1): p. 71-119.
21. S. Chakraborty, S. Sharma, P.K. Maiti, and Y. Krishnan, *The poly dA helix: a new structural motif for high performance DNA-based molecular switches*. Nucleic Acids Res, 2009. **37**(9): p. 2810-7.
22. A. B. J. A, L. J, and e. al., *Molecular Biology of the Cell, 4th Edition*. Garland Science, 2002. **18**(3).
23. S. Payne, *Introduction to RNA Viruses*, in *Viruses*. 2017. p. 97-105.
24. L. Zhang, A. Richards, M.I. Barrasa, S.H. Hughes, R.A. Young, and R. Jaenisch, *Reverse-transcribed SARS-CoV-2 RNA can integrate into the genome of cultured human cells and can be expressed in patient-derived tissues*. Proc Natl Acad Sci U S A, 2021. **118**(21).
25. Z. Lu, Q.C. Zhang, B. Lee, R.A. Flynn, M.A. Smith, J.T. Robinson, C. Davidovich, A.R. Gooding, K.J. Goodrich, J.S. Mattick, J.P. Mesirov, T.R. Cech, and H.Y. Chang, *RNA Duplex Map in Living Cells Reveals Higher-Order Transcriptome Structure*. Cell, 2016. **165**(5): p. 1267-1279.
26. N.J. Caplen, *Gene therapy progress and prospects. Downregulating gene expression: the impact of RNA interference*. Gene Ther, 2004. **11**(16): p. 1241-8.
27. C.W. Carter and J. Kraut, *A Proposed Model for Interaction of Polypeptides with RNA*. Proceedings of the National Academy of Sciences, 1974. **71**(2): p. 283.
28. B. Rauzan, E. McMichael, R. Cave, L.R. Sevcik, K. Ostrosky, E. Whitman, R. Stegemann, A.L. Sinclair, M.J. Serra, and A.A. Deckert, *Kinetics and thermodynamics of DNA, RNA, and hybrid duplex formation*. Biochemistry, 2013. **52**(5): p. 765-72.
29. T.E. Cheatham and P.A. Kollman, *Molecular Dynamics Simulations Highlight the Structural Differences among DNA:DNA, RNA:RNA, and DNA:RNA Hybrid Duplexes*. Journal of the American Chemical Society, 1997. **119**(21): p. 4805-4825.
30. J.I. Gyi, A.N. Lane, G.L. Conn, and T. Brown, *The orientation and dynamics of the C2'-OH and hydration of RNA and DNA-RNA hybrids*. Nucleic Acids Research, 1998. **26**(13): p. 3104-3110.
31. A. Bansal, M. Prasad, K. Roy, and S. Kukreti, *A short GC-rich palindrome of human mannose receptor gene coding region displays a conformational switch*. Biopolymers, 2012. **97**(12): p. 950-962.
32. H. Drew, T. Takano, S. Tanaka, K. Itakura, and R.E. Dickerson, *High-salt d(CpGpCpG), a left-handed Z' DNA double helix*. Nature, 1980. **286**(5773): p. 567-573.
33. A. Arcella, G. Portella, M.L. Ruiz, R. Eritja, M. Vilaseca, V. Gabelica, and M. Orozco, *Structure of triplex DNA in the gas phase*. J Am Chem Soc, 2012. **134**(15): p. 6596-606.
34. M.D. Frank-Kamenetskii and S.M. Mirkin, *TRIPLEX DNA STRUCTURES*. Annual Review of Biochemistry, 1995. **64**(1): p. 65-95.
35. J.L. Huppert, *Four-stranded nucleic acids: structure, function and targeting of G-quadruplexes*. Chem Soc Rev, 2008. **37**(7): p. 1375-84.

36. J.L. Mergny and D. Sen, *DNA Quadruple Helices in Nanotechnology*. Chem Rev, 2019. **119**(10): p. 6290-6325.
37. C. Chaput John and C. Switzer, *A DNA pentaplex incorporating nucleobase quintets*. Proceedings of the National Academy of Sciences, 1999. **96**(19): p. 10614-10619.
38. M. Gellert, M.N. Lipsett, and D.R. Davies, *Helix formation by guanylic acid*. Proc Natl Acad Sci U S A, 1962. **48**(12): p. 2013-8.
39. J.R. Williamson, M.K. Raghuraman, and T.R. Cech, *Monovalent cation-induced structure of telomeric DNA: the G-quartet model*. Cell, 1989. **59**(5): p. 871-80.
40. E. Largy, J.-L. Mergny, and V. Gabelica, *Role of Alkali Metal Ions in G-Quadruplex Nucleic Acid Structure and Stability*, in *The Alkali Metal Ions: Their Role for Life*, A. Sigel, H. Sigel, and R.K.O. Sigel, Editors. 2016, Springer International Publishing: Cham. p. 203-258.
41. F.R. Winnerdy and A.T. Phan, *Chapter Two - Quadruplex structure and diversity*, in *Annual Reports in Medicinal Chemistry*, S. Neidle, Editor. 2020, Academic Press. p. 45-73.
42. D.-H. Zhang, T. Fujimoto, S. Saxena, H.-Q. Yu, D. Miyoshi, and N. Sugimoto, *Monomorphic RNA G-Quadruplex and Polymorphic DNA G-Quadruplex Structures Responding to Cellular Environmental Factors*. Biochemistry, 2010. **49**(21): p. 4554-4563.
43. M.M. Fay, S.M. Lyons, and P. Ivanov, *RNA G-Quadruplexes in Biology: Principles and Molecular Mechanisms*. J Mol Biol, 2017. **429**(14): p. 2127-2147.
44. L. Lacroix, A. Seosse, and J.L. Mergny, *Fluorescence-based duplex-quadruplex competition test to screen for telomerase RNA quadruplex ligands*. Nucleic Acids Res, 2011. **39**(4): p. e21.
45. J.-L. Mergny and J.-C. Maurizot, *Fluorescence Resonance Energy Transfer as a Probe for G-Quartet Formation by a Telomeric Repeat*. ChemBioChem, 2001. **2**(2): p. 124-132.
46. E. Largy, A. Marchand, S. Amrane, V. Gabelica, and J.L. Mergny, *Quadruplex Turncoats: Cation-Dependent Folding and Stability of Quadruplex-DNA Double Switches*. J Am Chem Soc, 2016. **138**(8): p. 2780-92.
47. D. Miyoshi, A. Nakao, and N. Sugimoto, *Molecular Crowding Regulates the Structural Switch of the DNA G-Quadruplex*. Biochemistry, 2002. **41**(50): p. 15017-15024.
48. M. Cheng, J. Chen, H. Ju, J. Zhou, and J.L. Mergny, *Drivers of i-DNA Formation in a Variety of Environments Revealed by Four-Dimensional UV Melting and Annealing*. J Am Chem Soc, 2021. **143**(20): p. 7792-7807.
49. R. Buscaglia, M.C. Miller, W.L. Dean, R.D. Gray, A.N. Lane, J.O. Trent, and J.B. Chaires, *Polyethylene glycol binding alters human telomere G-quadruplex structure by conformational selection*. Nucleic Acids Res, 2013. **41**(16): p. 7934-46.
50. A. Guedin, J. Gros, P. Alberti, and J.L. Mergny, *How long is too long? Effects of loop size on G-quadruplex stability*. Nucleic Acids Res, 2010. **38**(21): p. 7858-68.
51. S. Amrane, M. Adrian, B. Heddi, A. Serero, A. Nicolas, J.L. Mergny, and A.T. Phan, *Formation of pearl-necklace monomorphic G-quadruplexes in the human CEB25 minisatellite*. J Am Chem Soc, 2012. **134**(13): p. 5807-16.
52. P. Hazel, J. Huppert, S. Balasubramanian, and S. Neidle, *Loop-Length-Dependent Folding of G-Quadruplexes*. Journal of the American Chemical Society, 2004. **126**(50): p. 16405-16415.
53. M. Cheng, Y. Cheng, J. Hao, G. Jia, J. Zhou, J.L. Mergny, and C. Li, *Loop permutation affects the topology and stability of G-quadruplexes*. Nucleic Acids Res, 2018. **46**(18): p. 9264-9275.

54. J. Chen, M. Cheng, G.F. Salgado, P. Stadlbauer, X. Zhang, S. Amrane, A. Guedin, F. He, J. Sponer, H. Ju, J.L. Mergny, and J. Zhou, *The beginning and the end: flanking nucleotides induce a parallel G-quadruplex topology*. *Nucleic Acids Res*, 2021. **49**(16): p. 9548-9559.
55. J.L. Leroy, K. Gehring, A. Kettani, and M. Guéron, *Acid multimers of oligodeoxycytidine strands: Stoichiometry, base-pair characterization, and proton exchange properties*. *Biochemistry*, 1993. **32**(23): p. 6019-6031.
56. K. Gehring, J.-L. Leroy, and M. Guéron, *A tetrameric DNA structure with protonated cytosine-cytosine base pairs*. *Nature*, 1993. **363**(6429): p. 561-565.
57. I. Berger, M. Egli, and A. Rich, *Inter-strand C-H...O hydrogen bonds stabilizing four-stranded intercalated molecules: stereoelectronic effects of O4' in cytosine-rich DNA*. *Proc Natl Acad Sci U S A*, 1996. **93**(22): p. 12116-21.
58. S. Benabou, A. Aviñó, R. Eritja, C. González, and R. Gargallo, *Fundamental aspects of the nucleic acid i-motif structures*. *RSC Adv.*, 2014. **4**(51): p. 26956-26980.
59. J.-L. Leroy, K. Snoussi, and M. Guéron, *Investigation of the energetics of C-H...O hydrogen bonds in the DNA i-motif via the equilibrium between alternative intercalation topologies*. *Magnetic Resonance in Chemistry*, 2001. **39**(S1): p. S171-S176.
60. I. Goncharova, *Ag(I)-mediated homo and hetero pairs of guanosine and cytidine: monitoring by circular dichroism spectroscopy*. *Spectrochim Acta A Mol Biomol Spectrosc*, 2014. **118**: p. 221-7.
61. J.-L. Mergny, L. Lacroix, X. Han, J.-L. Leroy, and C. Helene, *Intramolecular Folding of Pyrimidine Oligodeoxynucleotides into an i-DNA Motif*. *Journal of the American Chemical Society*, 1995. **117**(35): p. 8887-8898.
62. E.P. Wright, J.L. Huppert, and Z.A.E. Waller, *Identification of multiple genomic DNA sequences which form i-motif structures at neutral pH*. *Nucleic Acids Res*, 2017. **45**(6): p. 2951-2959.
63. M. Canalia and J.-L. Leroy, *[5mCCTCTCTCC]4: An i-Motif Tetramer with Intercalated T•T Pairs*. *Journal of the American Chemical Society*, 2009. **131**(36): p. 12870-12871.
64. M. Garavis, N. Escaja, V. Gabelica, A. Villasante, and C. Gonzalez, *Centromeric Alpha-Satellite DNA Adopts Dimeric i-Motif Structures Capped by AT Hoogsteen Base Pairs*. *Chemistry*, 2015. **21**(27): p. 9816-24.
65. J. Cui, P. Waltman, V.H. Le, and E.A. Lewis, *The effect of molecular crowding on the stability of human c-MYC promoter sequence I-motif at neutral pH*. *Molecules*, 2013. **18**(10): p. 12751-67.
66. J.H. Thorpe, S.C. Teixeira, B.C. Gale, and C.J. Cardin, *Crystal structure of the complementary quadruplex formed by d(GCATGCT) at atomic resolution*. *Nucleic Acids Res*, 2003. **31**(3): p. 844-9.
67. N. Zhang, A. Gorin, A. Majumdar, A. Kettani, N. Chernichenko, E. Skripkin, and D.J. Patel, *Dimeric DNA quadruplex containing major groove-aligned A-T-A-T and G-C-G-C tetrads stabilized by inter-subunit Watson-Crick A-T and G-C pairs*. *J Mol Biol*, 2001. **312**(5): p. 1073-88.
68. A. Kettani, S. Bouaziz, A. Gorin, H. Zhao, R.A. Jones, and D.J. Patel, *Solution structure of a Na cation stabilized DNA quadruplex containing G•G•G•G and G•C•G•C tetrads formed by G-G-G-C*

- repeats observed in adeno-associated viral DNA* Edited by I. Tinoco. *Journal of Molecular Biology*, 1998. **282**(3): p. 619-636.
69. N. Escaja, J.L. Gelpí, M. Orozco, M. Rico, E. Pedroso, and C. González, *Four-Stranded DNA Structure Stabilized by a Novel G:C:A:T Tetrad*. *Journal of the American Chemical Society*, 2003. **125**(19): p. 5654-5662.
 70. N. Maizels and L.T. Gray, *The G4 genome*. *PLoS Genet*, 2013. **9**(4): p. e1003468.
 71. D. Sen and W. Gilbert, *Formation of parallel four-stranded complexes by guanine-rich motifs in DNA and its implications for meiosis*. *Nature*, 1988. **334**(6180): p. 364-366.
 72. W.I. Sundquist and A. Klug, *Telomeric DNA dimerizes by formation of guanine tetrads between hairpin loops*. *Nature*, 1989. **342**(6251): p. 825-829.
 73. R.K. Moyzis, J.M. Buckingham, L.S. Cram, M. Dani, L.L. Deaven, M.D. Jones, J. Meyne, R.L. Ratliff, and J.R. Wu, *A highly conserved repetitive DNA sequence, (TTAGGG)_n, present at the telomeres of human chromosomes*. *Proceedings of the National Academy of Sciences*, 1988. **85**(18): p. 6622.
 74. A.K. Todd, M. Johnston, and S. Neidle, *Highly prevalent putative quadruplex sequence motifs in human DNA*. *Nucleic Acids Res*, 2005. **33**(9): p. 2901-7.
 75. J.L. Huppert and S. Balasubramanian, *Prevalence of quadruplexes in the human genome*. *Nucleic Acids Res*, 2005. **33**(9): p. 2908-16.
 76. W. Zhou, K. Suntharalingam, N.J. Brand, P.J. Barton, R. Vilar, and L. Ying, *Possible regulatory roles of promoter G-quadruplexes in cardiac function-related genes - human TnIc as a model*. *PLoS One*, 2013. **8**(1): p. e53137.
 77. E. Besnard, A. Babled, L. Lapasset, O. Milhavet, H. Parrinello, C. Dantec, J.M. Marin, and J.M. Lemaitre, *Unraveling cell type-specific and reprogrammable human replication origin signatures associated with G-quadruplex consensus motifs*. *Nat Struct Mol Biol*, 2012. **19**(8): p. 837-44.
 78. S.K. Bharti, J.A. Sommers, J. Zhou, D.L. Kaplan, J.N. Spelbrink, J.L. Mergny, and R.M. Brosh, Jr., *DNA sequences proximal to human mitochondrial DNA deletion breakpoints prevalent in human disease form G-quadruplexes, a class of DNA structures inefficiently unwound by the mitochondrial replicative Twinkle helicase*. *J Biol Chem*, 2014. **289**(43): p. 29975-93.
 79. V. Brazda, J. Lysek, M. Bartas, and M. Fojta, *Complex Analyses of Short Inverted Repeats in All Sequenced Chloroplast DNAs*. *Biomed Res Int*, 2018. **2018**: p. 1097018.
 80. S. Millevoi, H. Moine, and S. Vagner, *G-quadruplexes in RNA biology*. *Wiley Interdiscip Rev RNA*, 2012. **3**(4): p. 495-507.
 81. G.G. Jayaraj, S. Pandey, V. Scaria, and S. Maiti, *Potential G-quadruplexes in the human long non-coding transcriptome*. *RNA Biol*, 2012. **9**(1): p. 81-6.
 82. J. Song, J.P. Perreault, I. Topisirovic, and S. Richard, *RNA G-quadruplexes and their potential regulatory roles in translation*. *Translation (Austin)*, 2016. **4**(2): p. e1244031.
 83. S. Mestre-Fos, C. Ito, C.M. Moore, A.R. Reddi, and L.D. Williams, *Human ribosomal G-quadruplexes regulate heme bioavailability*. *J Biol Chem*, 2020. **295**(44): p. 14855-14865.
 84. F. Wu, K. Niu, Y. Cui, C. Li, M. Lyu, Y. Ren, Y. Chen, H. Deng, L. Huang, S. Zheng, L. Liu, J. Wang, Q. Song, H. Xiang, and Q. Feng, *Genome-wide analysis of DNA G-quadruplex motifs across 37 species provides insights into G4 evolution*. *Commun Biol*, 2021. **4**(1): p. 98.

85. N. Kosiol, S. Juranek, P. Brossart, A. Heine, and K. Paeschke, *G-quadruplexes: a promising target for cancer therapy*. *Mol Cancer*, 2021. **20**(1): p. 40.
86. J.A. Capra, K. Paeschke, M. Singh, and V.A. Zakian, *G-quadruplex DNA sequences are evolutionarily conserved and associated with distinct genomic features in Saccharomyces cerevisiae*. *PLoS Comput Biol*, 2010. **6**(7): p. e1000861.
87. P. Rawal, V.B. Kummarasetti, J. Ravindran, N. Kumar, K. Halder, R. Sharma, M. Mukerji, S.K. Das, and S. Chowdhury, *Genome-wide prediction of G4 DNA as regulatory motifs: role in Escherichia coli global regulation*. *Genome Res*, 2006. **16**(5): p. 644-55.
88. J.D. Beaudoin, R. Jodoin, and J.P. Perreault, *New scoring system to identify RNA G-quadruplex folding*. *Nucleic Acids Res*, 2014. **42**(2): p. 1209-23.
89. A. Bedrat, L. Lacroix, and J.L. Mergny, *Re-evaluation of G-quadruplex propensity with G4Hunter*. *Nucleic Acids Res*, 2016. **44**(4): p. 1746-59.
90. V. Brazda, J. Kolomaznik, J. Lysek, M. Bartas, M. Fojta, J. Stastny, and J.L. Mergny, *G4Hunter web application: a web server for G-quadruplex prediction*. *Bioinformatics*, 2019. **35**(18): p. 3493-3495.
91. V. Brazda, O. Porubiakova, A. Cantara, N. Bohalova, J. Coufal, M. Bartas, M. Fojta, and J.L. Mergny, *G-quadruplexes in H1N1 influenza genomes*. *BMC Genomics*, 2021. **22**(1): p. 77.
92. V. Brazda, Y. Luo, M. Bartas, P. Kaura, O. Porubiakova, J. Stastny, P. Pecinka, D. Verga, V. Da Cunha, T.S. Takahashi, P. Forterre, H. Myllykallio, M. Fojta, and J.L. Mergny, *G-Quadruplexes in the Archaea Domain*. *Biomolecules*, 2020. **10**(9).
93. A. Cantara, Y. Luo, M. Dobrovolna, N. Bohalova, M. Fojta, D. Verga, L. Guittat, A. Cucchiari, S. Savrimoutou, C. Haberli, J. Guillon, J. Keiser, V. Brazda, and J.L. Mergny, *G-quadruplexes in helminth parasites*. *Nucleic Acids Res*, 2022. **50**(5): p. 2719-2735.
94. R. Hansel-Hertsch, J. Spiegel, G. Marsico, D. Tannahill, and S. Balasubramanian, *Genome-wide mapping of endogenous G-quadruplex DNA structures by chromatin immunoprecipitation and high-throughput sequencing*. *Nat Protoc*, 2018. **13**(3): p. 551-564.
95. J.L. Huppert and S. Balasubramanian, *G-quadruplexes in promoters throughout the human genome*. *Nucleic Acids Res*, 2007. **35**(2): p. 406-13.
96. J. Eddy and N. Maizels, *Gene function correlates with potential for G4 DNA formation in the human genome*. *Nucleic Acids Res*, 2006. **34**(14): p. 3887-96.
97. J.Y. Gong, C.J. Wen, M.L. Tang, R.F. Duan, J.N. Chen, J.Y. Zhang, K.W. Zheng, Y.D. He, Y.H. Hao, Q. Yu, S.P. Ren, and Z. Tan, *G-quadruplex structural variations in human genome associated with single-nucleotide variations and their impact on gene activity*. *Proc Natl Acad Sci U S A*, 2021. **118**(21).
98. R.C. Monsen, L. DeLeeuw, W.L. Dean, R.D. Gray, T.M. Sabo, S. Chakravarthy, J.B. Chaires, and J.O. Trent, *The hTERT core promoter forms three parallel G-quadruplexes*. *Nucleic Acids Res*, 2020. **48**(10): p. 5720-5734.
99. A. Siddiqui-Jain, C.L. Grand, D.J. Bearss, and L.H. Hurley, *Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress c-myc transcription*. *Proceedings of the National Academy of Sciences*, 2002. **99**(18): p. 11593.

100. A.T. Phan, V. Kuryavyi, H.Y. Gaw, and D.J. Patel, *Small-molecule interaction with a five-guanine-tract G-quadruplex structure from the human MYC promoter*. *Nature Chemical Biology*, 2005. **1**(3): p. 167-173.
101. A.T. Phan, Y.S. Modi, and D.J. Patel, *Propeller-type parallel-stranded G-quadruplexes in the human c-myc promoter*. *J Am Chem Soc*, 2004. **126**(28): p. 8710-6.
102. A. Kerkour, J. Marquevielle, S. Ivashchenko, L.A. Yatsunyk, J.L. Mergny, and G.F. Salgado, *High-resolution three-dimensional NMR structure of the KRAS proto-oncogene promoter reveals key features of a G-quadruplex involved in transcriptional regulation*. *J Biol Chem*, 2017. **292**(19): p. 8082-8091.
103. S. Cogoi and L.E. Xodo, *G-quadruplex formation within the promoter of the KRAS proto-oncogene and its effect on transcription*. *Nucleic Acids Res*, 2006. **34**(9): p. 2536-49.
104. A. Membrino, S. Cogoi, E.B. Pedersen, and L.E. Xodo, *G4-DNA formation in the HRAS promoter and rational design of decoy oligonucleotides for cancer therapy*. *PLoS One*, 2011. **6**(9): p. e24421.
105. S.L. Palumbo, S.W. Ebbinghaus, and L.H. Hurley, *Formation of a Unique End-to-End Stacked Pair of G-Quadruplexes in the hTERT Core Promoter with Implications for Inhibition of Telomerase by G-Quadruplex-Interactive Ligands*. *Journal of the American Chemical Society*, 2009. **131**(31): p. 10878-10891.
106. S. Cogoi, A.E. Shchekotikhin, and L.E. Xodo, *HRAS is silenced by two neighboring G-quadruplexes and activated by MAZ, a zinc-finger transcription factor with DNA unfolding property*. *Nucleic Acids Res*, 2014. **42**(13): p. 8379-88.
107. D. Sun and L.H. Hurley, *The Importance of Negative Superhelicity in Inducing the Formation of G-Quadruplex and i-Motif Structures in the c-Myc Promoter: Implications for Drug Targeting and Control of Gene Expression*. *Journal of Medicinal Chemistry*, 2009. **52**(9): p. 2863-2874.
108. E. Tosoni, I. Frasson, M. Scalabrin, R. Perrone, E. Butovskaya, M. Nadai, G. Palu, D. Fabris, and S.N. Richter, *Nucleolin stabilizes G-quadruplex structures folded by the LTR promoter and silences HIV-1 viral transcription*. *Nucleic Acids Res*, 2015. **43**(18): p. 8884-97.
109. R. Perrone, M. Nadai, I. Frasson, J.A. Poe, E. Butovskaya, T.E. Smithgall, M. Palumbo, G. Palu, and S.N. Richter, *A dynamic G-quadruplex region regulates the HIV-1 long terminal repeat promoter*. *J Med Chem*, 2013. **56**(16): p. 6521-30.
110. S. Cogoi, M. Paramasivam, A. Membrino, K.K. Yokoyama, and L.E. Xodo, *The KRAS promoter responds to Myc-associated zinc finger and poly(ADP-ribose) polymerase 1 proteins, which recognize a critical quadruplex-forming GA-element*. *J Biol Chem*, 2010. **285**(29): p. 22003-16.
111. J. Eddy and N. Maizels, *Conserved elements with potential to form polymorphic G-quadruplex structures in the first intron of human genes*. *Nucleic Acids Res*, 2008. **36**(4): p. 1321-33.
112. C. Weldon, J.G. Dacanay, V. Gokhale, P.V.L. Boddupally, I. Behm-Ansmant, G.A. Burley, C. Branlant, L.H. Hurley, C. Dominguez, and I.C. Eperon, *Specific G-quadruplex ligands modulate the alternative splicing of Bcl-X*. *Nucleic Acids Res*, 2018. **46**(2): p. 886-896.
113. R.V. Brown, F.L. Danford, V. Gokhale, L.H. Hurley, and T.A. Brooks, *Demonstration that drug-targeted down-regulation of MYC in non-Hodgkins lymphoma is directly mediated through the promoter G-quadruplex*. *J Biol Chem*, 2011. **286**(47): p. 41018-27.

114. J.L. Huppert, A. Bugaut, S. Kumari, and S. Balasubramanian, *G-quadruplexes: the beginning and end of UTRs*. *Nucleic Acids Res*, 2008. **36**(19): p. 6260-8.
115. D.S.M. Lee, L.R. Ghanem, and Y. Barash, *Integrative analysis reveals RNA G-quadruplexes in UTRs are selectively constrained and enriched for functional associations*. *Nat Commun*, 2020. **11**(1): p. 527.
116. B.P. Belotserkovskii, S. Tornaletti, A.D. D'Souza, and P.C. Hanawalt, *R-loop generation during transcription: Formation, processing and cellular outcomes*. *DNA Repair (Amst)*, 2018. **71**: p. 69-81.
117. J. Tan and L. Lan, *The DNA secondary structures at telomeres and genome instability*. *Cell Biosci*, 2020. **10**: p. 47.
118. M.L. Duquette, P. Handa, J.A. Vincent, A.F. Taylor, and N. Maizels, *Intracellular transcription of G-rich DNAs induces formation of G-loops, novel structures containing G4 DNA*. *Genes Dev*, 2004. **18**(13): p. 1618-29.
119. P. Kotsantis, S. Segura-Bayona, P. Margalef, P. Marzec, P. Ruis, G. Hewitt, R. Bellelli, H. Patel, R. Goldstone, A.R. Poetsch, and S.J. Boulton, *RTEL1 Regulates G4/R-Loops to Avert Replication-Transcription Collisions*. *Cell Rep*, 2020. **33**(12): p. 108546.
120. A. De Magis, S.G. 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*. *Proc Natl Acad Sci U S A*, 2019. **116**(3): p. 816-825.
121. H. Xu, M. Di Antonio, S. McKinney, V. Mathew, B. Ho, N.J. O'Neil, N.D. Santos, J. Silvester, V. Wei, J. Garcia, F. Kabeer, D. Lai, P. Soriano, J. Banath, D.S. Chiu, D. Yap, D.D. Le, F.B. Ye, A. Zhang, K. Thu, J. Soong, S.C. Lin, A.H. Tsai, T. Osako, T. Algara, D.N. Saunders, J. Wong, J. Xian, M.B. Bally, J.D. Brenton, G.W. Brown, S.P. Shah, D. Cescon, T.W. Mak, C. Caldas, P.C. Stirling, P. Hieter, S. Balasubramanian, and S. Aparicio, *CX-5461 is a DNA G-quadruplex stabilizer with selective lethality in BRCA1/2 deficient tumours*. *Nat Commun*, 2017. **8**: p. 14432.
122. C.Y. Lee, C. Mc Nerney, K. Ma, W. Zhao, A. Wang, and S. Myong, *R-loop induced G-quadruplex in non-template promotes transcription by successive R-loop formation*. *Nat Commun*, 2020. **11**(1): p. 3392.
123. M. Jara-Espejo and S.R. Line, *DNA G-quadruplex stability, position and chromatin accessibility are associated with CpG island methylation*. *FEBS J*, 2020. **287**(3): p. 483-495.
124. A.L. Valton, V. Hassan-Zadeh, I. Lema, N. Boggetto, P. Alberti, C. Saintome, J.F. Riou, and M.N. Prioleau, *G4 motifs affect origin positioning and efficiency in two vertebrate replicators*. *EMBO J*, 2014. **33**(7): p. 732-46.
125. L.J. Reha-Krantz, *Okazaki Fragment*, in *Brenner's Encyclopedia of Genetics (Second Edition)*, S. Maloy and K. Hughes, Editors. 2013, Academic Press: San Diego. p. 158-160.
126. I. Cheung, M. Schertzer, A. Rose, and P.M. Lansdorp, *Disruption of dog-1 in Caenorhabditis elegans triggers deletions upstream of guanine-rich DNA*. *Nature Genetics*, 2002. **31**(4): p. 405-409.
127. J. Lopes, A. Piazza, R. Bermejo, B. Kriegsman, A. Colosio, M.P. Teulade-Fichou, M. Foiani, and A. Nicolas, *G-quadruplex-induced instability during leading-strand replication*. *EMBO J*, 2011. **30**(19): p. 4033-46.

128. D. Schiavone, G. Guilbaud, P. Murat, C. Papadopoulou, P. Sarkies, M.N. Prioleau, S. Balasubramanian, and J.E. Sale, *Determinants of G quadruplex-induced epigenetic instability in REV1-deficient cells*. EMBO J, 2014. **33**(21): p. 2507-20.
129. I. Tinoco, *Hypochromism in Polynucleotides1*. Journal of the American Chemical Society, 1960. **82**(18): p. 4785-4790.
130. G.C.K. Roberts, *Encyclopedia of Biophysics*. 2013.
131. P.V. Scaria, S.J. Shire, and R.H. Shafer, *Quadruplex structure of d(G3T4G3) stabilized by K+ or Na+ is an asymmetric hairpin dimer*. Proceedings of the National Academy of Sciences, 1992. **89**(21): p. 10336.
132. J.L. Mergny, J. Li, L. Lacroix, S. Amrane, and J.B. Chaires, *Thermal difference spectra: a specific signature for nucleic acid structures*. Nucleic Acids Res, 2005. **33**(16): p. e138.
133. A. Guedin, L.Y. Lin, S. Armane, L. Lacroix, J.L. Mergny, S. Thore, and L.A. Yatsunyk, *Quadruplexes in 'Dicty': crystal structure of a four-quartet G-quadruplex formed by G-rich motif found in the Dictyostelium discoideum genome*. Nucleic Acids Res, 2018. **46**(10): p. 5297-5307.
134. J.-L. Mergny and L. Lacroix, *Analysis of Thermal Melting Curves*. Oligonucleotides, 2003. **13**(6): p. 515-537.
135. J.L. Mergny and L. Lacroix, *UV Melting of G-Quadruplexes*. Curr Protoc Nucleic Acid Chem, 2009. **Chapter 17**: p. Unit 17 1.
136. S. Paramasivan, I. Rujan, and P.H. Bolton, *Circular dichroism of quadruplex DNAs: applications to structure, cation effects and ligand binding*. Methods, 2007. **43**(4): p. 324-31.
137. S. Masiero, R. Trotta, S. Pieraccini, S. De Tito, R. Perone, A. Randazzo, and G.P. Spada, *A non-empirical chromophoric interpretation of CD spectra of DNA G-quadruplex structures*. Org Biomol Chem, 2010. **8**(12): p. 2683-92.
138. J.-D. Wen and D.M. Gray, *The Ff Gene 5 Single-Stranded DNA-Binding Protein Binds to the Transiently Folded Form of an Intramolecular G-Quadruplex*. Biochemistry, 2002. **41**(38): p. 11438-11448.
139. D.M. Gray, J.D. Wen, C.W. Gray, R. Repges, C. Repges, G. Raabe, and J. Fleischhauer, *Measured and calculated CD spectra of G-quartets stacked with the same or opposite polarities*. Chirality, 2008. **20**(3-4): p. 431-40.
140. R.C. Monsen, L.W. DeLeeuw, W.L. Dean, R.D. Gray, S. Chakravarthy, J.B. Hopkins, J.B. Chaires, and J.O. Trent, *Long promoter sequences form higher-order G-quadruplexes: an integrative structural biology study of c-Myc, k-Ras and c-Kit promoter sequences*. Nucleic Acids Res, 2022. **50**(7): p. 4127-4147.
141. P. Tothova, P. Krafcikova, and V. Viglasky, *Formation of highly ordered multimers in G-quadruplexes*. Biochemistry, 2014. **53**(45): p. 7013-27.
142. R. Del Villar-Guerra, J.O. Trent, and J.B. Chaires, *G-Quadruplex Secondary Structure Obtained from Circular Dichroism Spectroscopy*. Angew Chem Int Ed Engl, 2018. **57**(24): p. 7171-7175.
143. A. Virgilio, V. Esposito, A. Randazzo, L. Mayol, and A. Galeone, *8-methyl-2'-deoxyguanosine incorporation into parallel DNA quadruplex structures*. Nucleic Acids Res, 2005. **33**(19): p. 6188-95.
144. J. Kypr, I. Kejnovska, D. Renciuik, and M. Vorlickova, *Circular dichroism and conformational polymorphism of DNA*. Nucleic Acids Res, 2009. **37**(6): p. 1713-25.

145. M. Adrian, B. Heddi, and A.T. Phan, *NMR spectroscopy of G-quadruplexes*. *Methods*, 2012. **57**(1): p. 11-24.
146. S. Dzatko, M. Krafcikova, R. Hansel-Hertsch, T. Fessl, R. Fiala, T. Loja, D. Krafcik, J.L. Mergny, S. Foldynova-Trantirkova, and L. Trantirek, *Evaluation of the Stability of DNA i-Motifs in the Nuclei of Living Mammalian Cells*. *Angew Chem Int Ed Engl*, 2018. **57**(8): p. 2165-2169.
147. K.D. Berger, S.D. Kennedy, and D.H. Turner, *Nuclear Magnetic Resonance Reveals That GU Base Pairs Flanking Internal Loops Can Adopt Diverse Structures*. *Biochemistry*, 2019. **58**(8): p. 1094-1108.
148. S. Kolesnikova, M. Hubalek, L. Bednarova, J. Cvacka, and E.A. Curtis, *Multimerization rules for G-quadruplexes*. *Nucleic Acids Res*, 2017. **45**(15): p. 8684-8696.
149. J.S. Smith and F.B. Johnson, *Isolation of G-quadruplex DNA using NMM-sepharose affinity chromatography*. *Methods Mol Biol*, 2010. **608**: p. 207-21.
150. A. Renaud de la Faverie, A. Guedin, A. Bedrat, L.A. Yatsunyk, and J.L. Mergny, *Thioflavin T as a fluorescence light-up probe for G4 formation*. *Nucleic Acids Res*, 2014. **42**(8): p. e65.
151. M. Zuffo, A. Guedin, E.D. Leriche, F. Doria, V. Pirola, V. Gabelica, J.L. Mergny, and M. Freccero, *More is not always better: finding the right trade-off between affinity and selectivity of a G-quadruplex ligand*. *Nucleic Acids Res*, 2018. **46**(19): p. e115.
152. M. Zuffo, A. Gandolfini, B. Heddi, and A. Granzhan, *Harnessing intrinsic fluorescence for typing of secondary structures of DNA*. *Nucleic Acids Res*, 2020. **48**(11): p. e61.
153. T. Gustavsson and D. Markovitsi, *Fundamentals of the Intrinsic DNA Fluorescence*. *Acc Chem Res*, 2021. **54**(5): p. 1226-1235.
154. R. Improta, *Quantum mechanical calculations unveil the structure and properties of the absorbing and emitting excited electronic states of guanine quadruplex*. *Chemistry*, 2014. **20**(26): p. 8106-15.
155. A.J. Lawaetz and C.A. Stedmon, *Fluorescence Intensity Calibration Using the Raman Scatter Peak of Water*. *Applied Spectroscopy*, 2009. **63**(8): p. 936-940.
156. A.C. Bhasikuttan and J. Mohanty, *Targeting G-quadruplex structures with extrinsic fluorogenic dyes: promising fluorescence sensors*. *Chem Commun (Camb)*, 2015. **51**(36): p. 7581-97.
157. E. Largy, A. Granzhan, F. Hamon, D. Verga, and M.-P. Teulade-Fichou, *Visualizing the Quadruplex: From Fluorescent Ligands to Light-Up Probes*, in *Quadruplex Nucleic Acids*, J.B. Chaires and D. Graves, Editors. 2013, Springer Berlin Heidelberg: Berlin, Heidelberg. p. 111-177.
158. A. Yett, L.Y. Lin, D. Beseiso, J. Miao, and L.A. Yatsunyk, *N-methyl mesoporphyrin IX as a highly selective light-up probe for G-quadruplex DNA*. *J Porphyr Phthalocyanines*, 2019. **23**(11n12): p. 1195-1215.
159. M. Zuffo, F. Doria, S. Botti, G. Bergamaschi, and M. Freccero, *G-quadruplex fluorescence sensing by core-extended naphthalene diimides*. *Biochim Biophys Acta Gen Subj*, 2017. **1861**(5 Pt B): p. 1303-1311.
160. A.C. Bhasikuttan, J. Mohanty, and H. Pal, *Interaction of malachite green with guanine-rich single-stranded DNA: preferential binding to a G-quadruplex*. *Angew Chem Int Ed Engl*, 2007. **46**(48): p. 9305-7.

161. D.M. Kong, Y.E. Ma, J. Wu, and H.X. Shen, *Discrimination of G-quadruplexes from duplex and single-stranded DNAs with fluorescence and energy-transfer fluorescence spectra of crystal violet*. *Chemistry*, 2009. **15**(4): p. 901-9.
162. Y. Kataoka, H. Fujita, Y. Kasahara, T. Yoshihara, S. Tobita, and M. Kuwahara, *Minimal Thioflavin T Modifications Improve Visual Discrimination of Guanine-Quadruplex Topologies and Alter Compound-Induced Topological Structures*. *Analytical Chemistry*, 2014. **86**(24): p. 12078-12084.
163. X. Xie, M. Zuffo, M.P. Teulade-Fichou, and A. Granzhan, *Identification of optimal fluorescent probes for G-quadruplex nucleic acids through systematic exploration of mono- and distyryl dye libraries*. *Beilstein J Org Chem*, 2019. **15**: p. 1872-1889.
164. J.M. Nicoludis, S.P. Barrett, J.L. Mergny, and L.A. Yatsunyk, *Interaction of human telomeric DNA with N-methyl mesoporphyrin IX*. *Nucleic Acids Res*, 2012. **40**(12): p. 5432-47.
165. V. Gabelica, R. Maeda, T. Fujimoto, H. Yaku, T. Murashima, N. Sugimoto, and D. Miyoshi, *Multiple and Cooperative Binding of Fluorescence Light-up Probe Thioflavin T with Human Telomere DNA G-Quadruplex*. *Biochemistry*, 2013. **52**(33): p. 5620-5628.
166. E. Largy and J.L. Mergny, *Shape matters: size-exclusion HPLC for the study of nucleic acid structural polymorphism*. *Nucleic Acids Res*, 2014. **42**(19): p. e149.
167. M.M. Dailey, M.C. Miller, P.J. Bates, A.N. Lane, and J.O. Trent, *Resolution and characterization of the structural polymorphism of a single quadruplex-forming sequence*. *Nucleic Acids Res*, 2010. **38**(14): p. 4877-88.
168. W.L. Dean, R.D. Gray, L. DeLeeuw, R.C. Monsen, and J.B. Chaires, *Putting a New Spin of G-Quadruplex Structure and Binding by Analytical Ultracentrifugation*. *Methods Mol Biol*, 2019. **2035**: p. 87-103.
169. V. Gabelica, *Native Mass Spectrometry and Nucleic Acid G-Quadruplex Biophysics: Advancing Hand in Hand*. *Accounts of Chemical Research*, 2021. **54**(19): p. 3691-3699.
170. E. Largy, A. König, A. Ghosh, D. Ghosh, S. Benabou, F. Rosu, and V. Gabelica, *Mass Spectrometry of Nucleic Acid Noncovalent Complexes*. *Chemical Reviews*, 2022. **122**(8): p. 7720-7839.
171. S. Daly, F. Rosu, and V. Gabelica, *Mass-resolved electronic circular dichroism ion spectroscopy*. *Science*, 2020. **368**(6498): p. 1465-1468.
172. D. Sun and L.H. Hurley, *Biochemical techniques for the characterization of G-quadruplex structures: EMSA, DMS footprinting, and DNA polymerase stop assay*. *Methods Mol Biol*, 2010. **608**: p. 65-79.
173. D. Sun, B. Thompson, B.E. Cathers, M. Salazar, S.M. Kerwin, J.O. Trent, T.C. Jenkins, S. Neidle, and L.H. Hurley, *Inhibition of Human Telomerase by a G-Quadruplex-Interactive Compound*. *Journal of Medicinal Chemistry*, 1997. **40**(14): p. 2113-2116.
174. Q. Li, J.F. Xiang, Q.F. Yang, H.X. Sun, A.J. Guan, and Y.L. Tang, *G4LDB: a database for discovering and studying G-quadruplex ligands*. *Nucleic Acids Res*, 2013. **41**(Database issue): p. D1115-23.
175. S. Neidle, *Human telomeric G-quadruplex: the current status of telomeric G-quadruplexes as therapeutic targets in human cancer*. *FEBS J*, 2010. **277**(5): p. 1118-25.

176. D.J. Patel, A.T. Phan, and V. Kuryavyi, *Human telomere, oncogenic promoter and 5'-UTR G-quadruplexes: diverse higher order DNA and RNA targets for cancer therapeutics*. *Nucleic Acids Res*, 2007. **35**(22): p. 7429-55.
177. A. Ghosh, M. Trajkovski, M.P. Teulade-Fichou, V. Gabelica, and J. Plavec, *Phen-DC3 Induces Refolding of Human Telomeric DNA into a Chair-Type Antiparallel G-Quadruplex through Ligand Intercalation*. *Angew Chem Int Ed Engl*, 2022: p. e202207384.
178. A. Funke, B. Karg, J. Dickerhoff, D. Balke, S. Muller, and K. Weisz, *Ligand-Induced Dimerization of a Truncated Parallel MYC G-Quadruplex*. *Chembiochem*, 2018. **19**(5): p. 505-512.
179. S. Roy, A. Ali, M. Kamra, K. Muniyappa, and S. Bhattacharya, *Specific stabilization of promoter G-Quadruplex DNA by 2,6-disubstituted amidoanthracene-9,10-dione based dimeric distamycin analogues and their selective cancer cell cytotoxicity*. *Eur J Med Chem*, 2020. **195**: p. 112202.
180. R.T. Wheelhouse, D. Sun, H. Han, F.X. Han, and L.H. Hurley, *Cationic Porphyrins as Telomerase Inhibitors: the Interaction of Tetra-(N-methyl-4-pyridyl)porphine with Quadruplex DNA*. *Journal of the American Chemical Society*, 1998. **120**(13): p. 3261-3262.
181. E. Izbiccka, R.T. Wheelhouse, E. Raymond, K.K. Davidson, R.A. Lawrence, D. Sun, B.E. Windle, L.H. Hurley, and D.D. Von Hoff, *Effects of Cationic Porphyrins as G-Quadruplex Interactive Agents in Human Tumor Cells*. *Cancer Research*, 1999. **59**(3): p. 639.
182. J. Ren and J.B. Chaires, *Sequence and Structural Selectivity of Nucleic Acid Binding Ligands*. *Biochemistry*, 1999. **38**(49): p. 16067-16075.
183. Y. Li and D. Sen, *A catalytic DNA for porphyrin metallation*. *Nature Structural Biology*, 1996. **3**(9): p. 743-747.
184. L. Sabater, M.L. Nicolau-Travers, A. De Rache, E. Prado, J. Dejeu, O. Bombarde, J. Lacroix, P. Calsou, E. Defrancq, J.L. Mergny, D. Gomez, and G. Pratviel, *The nickel(II) complex of guanidinium phenyl porphyrin, a specific G-quadruplex ligand, targets telomeres and leads to POT1 mislocalization in culture cells*. *J Biol Inorg Chem*, 2015. **20**(4): p. 729-38.
185. I.M. Dixon, F. Lopez, A.M. Tejera, J.-P. Estève, M.A. Blasco, G. Pratviel, and B. Meunier, *A G-Quadruplex Ligand with 10000-Fold Selectivity over Duplex DNA*. *Journal of the American Chemical Society*, 2007. **129**(6): p. 1502-1503.
186. Q. Cao, Y. Li, E. Freisinger, P.Z. Qin, R.K.O. Sigel, and Z.-W. Mao, *G-quadruplex DNA targeted metal complexes acting as potential anticancer drugs*. *Inorganic Chemistry Frontiers*, 2017. **4**(1): p. 10-32.
187. R.J. Harrison, J. Cuesta, G. Chessari, M.A. Read, S.K. Basra, A.P. Reszka, J. Morrell, S.M. Gowan, C.M. Incles, F.A. Tanius, W.D. Wilson, L.R. Kelland, and S. Neidle, *Trisubstituted Acridine Derivatives as Potent and Selective Telomerase Inhibitors*. *Journal of Medicinal Chemistry*, 2003. **46**(21): p. 4463-4476.
188. A.M. Burger, F. Dai, C.M. Schultes, A.P. Reszka, M.J. Moore, J.A. Double, and S. Neidle, *The G-Quadruplex-Interactive Molecule BRACO-19 Inhibits Tumor Growth, Consistent with Telomere Targeting and Interference with Telomerase Function*. *Cancer Research*, 2005. **65**(4): p. 1489.

189. M. Gunaratnam, C. Green, J.B. Moreira, A.D. Moorhouse, L.R. Kelland, J.E. Moses, and S. Neidle, *G-quadruplex compounds and cis-platin act synergistically to inhibit cancer cell growth in vitro and in vivo*. *Biochem Pharmacol*, 2009. **78**(2): p. 115-22.
190. J. Carvalho, E. Pereira, J. Marquevielle, M.P.C. Campello, J.L. Mergny, A. Paulo, G.F. Salgado, J.A. Queiroz, and C. Cruz, *Fluorescent light-up acridine orange derivatives bind and stabilize KRAS-22RT G-quadruplex*. *Biochimie*, 2018. **144**: p. 144-152.
191. M.-Y. Kim, W. Duan, M. Gleason-Guzman, and L.H. Hurley, *Design, Synthesis, and Biological Evaluation of a Series of Fluoroquinoanthroxazines with Contrasting Dual Mechanisms of Action against Topoisomerase II and G-Quadruplexes*. *Journal of Medicinal Chemistry*, 2003. **46**(4): p. 571-583.
192. D. Drygin, A. Siddiqui-Jain, S. O'Brien, M. Schwaebe, A. Lin, J. Bliesath, C.B. Ho, C. Proffitt, K. Trent, J.P. Whitten, J.K. Lim, D. Von Hoff, K. Anderes, and W.G. Rice, *Anticancer activity of CX-3543: a direct inhibitor of rRNA biogenesis*. *Cancer Res*, 2009. **69**(19): p. 7653-61.
193. Y. Katsuda, S. Sato, L. Asano, Y. Morimura, T. Furuta, H. Sugiyama, M. Hagihara, and M. Uesugi, *A Small Molecule That Represses Translation of G-Quadruplex-Containing mRNA*. *J Am Chem Soc*, 2016. **138**(29): p. 9037-40.
194. R. Rodriguez, S. Müller, J.A. Yeoman, C. Trentesaux, J.-F. Riou, and S. Balasubramanian, *A Novel Small Molecule That Alters Shelterin Integrity and Triggers a DNA-Damage Response at Telomeres*. *Journal of the American Chemical Society*, 2008. **130**(47): p. 15758-15759.
195. J.F. Moruno-Manchon, P. Lejault, Y. Wang, B. McCauley, P. Honarpisheh, D.A. Morales Scheihing, S. Singh, W. Dang, N. Kim, A. Urayama, L. Zhu, D. Monchaud, L.D. McCullough, and A.S. Tsvetkov, *Small-molecule G-quadruplex stabilizers reveal a novel pathway of autophagy regulation in neurons*. *Elife*, 2020. **9**.
196. F. Wu, C. Liu, Y. Chen, S. Yang, J. Xu, R. Huang, X. Wang, M. Li, W. Liu, W. Mao, and X. Zhou, *Visualization of G-quadruplexes in gel and in live cells by a near-infrared fluorescent probe*. *Sensors and Actuators B: Chemical*, 2016. **236**: p. 268-275.
197. R.N. Das, E. Chevret, V. Desplat, S. Rubio, J.L. Mergny, and J. Guillon, *Design, Synthesis and Biological Evaluation of New Substituted Diquinolinyl-Pyridine Ligands as Anticancer Agents by Targeting G-Quadruplex*. *Molecules*, 2017. **23**(1).
198. G. Pennarun, C. Granotier, L.R. Gauthier, D. Gomez, F. Hoffschir, E. Mandine, J.-F. Riou, J.-L. Mergny, P. Mailliet, and F.D. Boussin, *Apoptosis related to telomere instability and cell cycle alterations in human glioma cells treated by new highly selective G-quadruplex ligands*. *Oncogene*, 2005. **24**(18): p. 2917-2928.
199. A. De Cian, E. DeLemos, J.-L. Mergny, M.-P. Teulade-Fichou, and D. Monchaud, *Highly Efficient G-Quadruplex Recognition by Bisquinolinium Compounds*. *Journal of the American Chemical Society*, 2007. **129**(7): p. 1856-1857.
200. R. Halder, J.F. Riou, M.P. Teulade-Fichou, T. Frickey, and J.S. Hartig, *Bisquinolinium compounds induce quadruplex-specific transcriptome changes in HeLa S3 cell lines*. *BMC Res Notes*, 2012. **5**: p. 138.
201. J. Lefebvre, C. Guetta, F. Poyer, F. Mahuteau-Betzer, and M.P. Teulade-Fichou, *Copper-Alkyne Complexation Responsible for the Nucleolar Localization of Quadruplex Nucleic Acid Drugs Labeled by Click Reactions*. *Angew Chem Int Ed Engl*, 2017. **56**(38): p. 11365-11369.

202. V. Pirola, M. Nadai, F. Doria, and S.N. Richter, *Naphthalene Diimides as Multimodal G-Quadruplex-Selective Ligands*. *Molecules*, 2019. **24**(3).
203. G.W. Collie, R. Promontorio, S.M. Hampel, M. Micco, S. Neidle, and G.N. Parkinson, *Structural Basis for Telomeric G-Quadruplex Targeting by Naphthalene Diimide Ligands*. *Journal of the American Chemical Society*, 2012. **134**(5): p. 2723-2731.
204. J.H. Guo, L.N. Zhu, D.M. Kong, and H.X. Shen, *Triphenylmethane dyes as fluorescent probes for G-quadruplex recognition*. *Talanta*, 2009. **80**(2): p. 607-13.
205. J. Flinders, S.C. DeFina, D.M. Brackett, C. Baugh, C. Wilson, and T. Dieckmann, *Recognition of planar and nonplanar ligands in the malachite green-RNA aptamer complex*. *Chembiochem*, 2004. **5**(1): p. 62-72.
206. D.-M. Kong, Y.-E. Ma, J.-H. Guo, W. Yang, and H.-X. Shen, *Fluorescent Sensor for Monitoring Structural Changes of G-Quadruplexes and Detection of Potassium Ion*. *Analytical Chemistry*, 2009. **81**(7): p. 2678-2684.
207. X.Y. Zhang, H.Q. Luo, and N.B. Li, *Crystal violet as an i-motif structure probe for reversible and label-free pH-driven electrochemical switch*. *Anal Biochem*, 2014. **455**: p. 55-9.
208. S. Zhang, H. Sun, H. Chen, Q. Li, A. Guan, L. Wang, Y. Shi, S. Xu, M. Liu, and Y. Tang, *Direct visualization of nucleolar G-quadruplexes in live cells by using a fluorescent light-up probe*. *Biochim Biophys Acta Gen Subj*, 2018. **1862**(5): p. 1101-1106.
209. L. Liu, Y. Shao, J. Peng, H. Liu, and L. Zhang, *Selective recognition of ds-DNA cavities by a molecular rotor: switched fluorescence of thioflavin T*. *Mol Biosyst*, 2013. **9**(10): p. 2512-9.
210. S. Liu, P. Peng, H. Wang, L. Shi, and T. Li, *Thioflavin T binds dimeric parallel-stranded GA-containing non-G-quadruplex DNAs: a general approach to lighting up double-stranded scaffolds*. *Nucleic Acids Res*, 2017. **45**(21): p. 12080-12089.
211. Y. Kataoka, H. Fujita, T. Endoh, N. Sugimoto, and M. Kuwahara, *Effects of Modifying Thioflavin T at the N(3)-Position on Its G4 Binding and Fluorescence Emission*. *Molecules*, 2020. **25**(21).
212. Y. Yan, J. Tan, T. Ou, Z. Huang, and L. Gu, *DNA G-quadruplex binders: a patent review*. *Expert Opinion on Therapeutic Patents* 2013. **23**(11): p. 1495-1509.
213. L. Martino, A. Virno, B. Pagano, A. Virgilio, S. Di Micco, A. Galeone, C. Giancola, G. Bifulco, L. Mayol, and A. Randazzo, *Structural and Thermodynamic Studies of the Interaction of Distamycin A with the Parallel Quadruplex Structure [d(TGGGGT)]₄*. *Journal of the American Chemical Society*, 2007. **129**(51): p. 16048-16056.
214. C. Rajput, R. Rutkaite, L. Swanson, I. Haq, and J.A. Thomas, *Dinuclear Monointercalating Rull Complexes That Display High Affinity Binding to Duplex and Quadruplex DNA*. *Chemistry – A European Journal*, 2006. **12**(17): p. 4611-4619.
215. P. Weisman-Shomer, E. Cohen, I. Hershco, S. Khateb, O. Wolfvovitz-Barchad, L.H. Hurley, and M. Fry, *The cationic porphyrin TMPyP4 destabilizes the tetraplex form of the fragile X syndrome expanded sequence d(CGG)_n*. *Nucleic Acids Res*, 2003. **31**(14): p. 3963-70.
216. J. Mitteau, P. Lejault, F. Wojciechowski, A. Joubert, J. Boudon, N. Desbois, C.P. Gros, R.H.E. Hudson, J.B. Boule, A. Granzhan, and D. Monchaud, *Identifying G-Quadruplex-DNA-Disrupting Small Molecules*. *J Am Chem Soc*, 2021. **143**(32): p. 12567-12577.

217. S. Christian, J. Pilch, M.E. Akerman, K. Porkka, P. Laakkonen, and E. Ruoslahti, *Nucleolin expressed at the cell surface is a marker of endothelial cells in angiogenic blood vessels*. J Cell Biol, 2003. **163**(4): p. 871-8.
218. D. Destouches, D. El Khoury, Y. Hamma-Kourbali, B. Krust, P. Albanese, P. Katsoris, G. Guichard, J.P. Briand, J. Courty, and A.G. Hovanessian, *Suppression of tumor growth and angiogenesis by a specific antagonist of the cell-surface expressed nucleolin*. PLoS One, 2008. **3**(6): p. e2518.
219. F. Morfoisse, F. Tatin, F. Hantelys, A. Adoue, A.C. Helfer, S. Cassant-Sourdy, F. Pujol, A. Gomez-Brouchet, L. Ligat, F. Lopez, S. Pyronnet, J. Courty, J. Guillermet-Guibert, S. Marzi, R.J. Schneider, A.C. Prats, and B.H. Garmy-Susini, *Nucleolin Promotes Heat Shock-Associated Translation of VEGF-D to Promote Tumor Lymphangiogenesis*. Cancer Res, 2016. **76**(15): p. 4394-405.
220. P.J. Bates, J.B. Kahlon, S.D. Thomas, J.O. Trent, and D.M. Miller, *Antiproliferative activity of G-rich oligonucleotides correlates with protein binding*. J Biol Chem, 1999. **274**(37): p. 26369-77.
221. J.E. Rosenberg, R.M. Bambury, E.M. Van Allen, H.A. Drabkin, P.N. Lara, Jr., A.L. Harzstark, N. Wagle, R.A. Figlin, G.W. Smith, L.A. Garraway, T. Choueiri, F. Erlandsson, and D.A. Laber, *A phase II trial of AS1411 (a novel nucleolin-targeted DNA aptamer) in metastatic renal cell carcinoma*. Investigational new drugs, 2014. **32**(1): p. 178-187.
222. T. Wang, Y. Luo, H. Lv, J. Wang, Y. Zhang, and R. Pei, *Aptamer-Based Erythrocyte-Derived Mimic Vesicles Loaded with siRNA and Doxorubicin for the Targeted Treatment of Multidrug-Resistant Tumors*. ACS Appl Mater Interfaces, 2019. **11**(49): p. 45455-45466.
223. K.N. Estep, T.J. Butler, J. Ding, and R.M. Brosh, *G4-Interacting DNA Helicases and Polymerases: Potential Therapeutic Targets*. Curr Med Chem, 2019. **26**(16): p. 2881-2897.
224. Y. Wu, K. Shin-ya, and R.M. Brosh, Jr., *FANCI helicase defective in Fanconia anemia and breast cancer unwinds G-quadruplex DNA to defend genomic stability*. Mol Cell Biol, 2008. **28**(12): p. 4116-28.
225. J.B. Budhathoki, S. Ray, V. Urban, P. Janscak, J.G. Yodh, and H. Balci, *RecQ-core of BLM unfolds telomeric G-quadruplex in the absence of ATP*. Nucleic Acids Res, 2014. **42**(18): p. 11528-45.
226. O. Mendoza, A. Bourdoncle, J.B. Boule, R.M. Brosh, Jr., and J.L. Mergny, *G-quadruplexes and helicases*. Nucleic Acids Res, 2016. **44**(5): p. 1989-2006.
227. G. Wu, Z. Xing, E.J. Tran, and D. Yang, *DDX5 helicase resolves G-quadruplex and is involved in MYC gene transcriptional activation*. Proc Natl Acad Sci U S A, 2019. **116**(41): p. 20453-20461.
228. E.P. Booy, M. Meier, N. Okun, S.K. Novakowski, S. Xiong, J. Stetefeld, and S.A. McKenna, *The RNA helicase RHAU (DHX36) unwinds a G4-quadruplex in human telomerase RNA and promotes the formation of the P1 helix template boundary*. Nucleic Acids Res, 2012. **40**(9): p. 4110-24.
229. K. Hiom, *FANCI: solving problems in DNA replication*. DNA Repair (Amst), 2010. **9**(3): p. 250-6.

230. N.M. Gueddouda, O. Mendoza, D. Gomez, A. Bourdoncle, and J.L. Mergny, *G-quadruplexes unfolding by RHAU helicase*. *Biochim Biophys Acta Gen Subj*, 2017. **1861**(5 Pt B): p. 1382-1388.
231. B.A. Brown, Y. Li, J.C. Brown, C.C. Hardin, J.F. Roberts, S.C. Pelsue, and L.D. Shultz, *Isolation and Characterization of a Monoclonal Anti-Quadruplex DNA Antibody from Autoimmune "Viable Motheaten" Mice*. *Biochemistry*, 1998. **37**(46): p. 16325-16337.
232. H. Fernando, R. Rodriguez, and S. Balasubramanian, *Selective Recognition of a DNA G-Quadruplex by an Engineered Antibody*. *Biochemistry*, 2008. **47**(36): p. 9365-9371.
233. E.Y. Lam, D. Beraldi, D. Tannahill, and S. Balasubramanian, *G-quadruplex structures are stable and detectable in human genomic DNA*. *Nat Commun*, 2013. **4**: p. 1796.
234. K.W. Zheng, J.Y. Zhang, Y.D. He, J.Y. Gong, C.J. Wen, J.N. Chen, Y.H. Hao, Y. Zhao, and Z. Tan, *Detection of genomic G-quadruplexes in living cells using a small artificial protein*. *Nucleic Acids Res*, 2020. **48**(20): p. 11706-11720.
235. H.G. Kazemier, K. Paeschke, and P.M. Lansdorp, *Guanine quadruplex monoclonal antibody 1H6 cross-reacts with restrained thymidine-rich single stranded DNA*. *Nucleic Acids Res*, 2017. **45**(10): p. 5913-5919.
236. G. Biffi, D. Tannahill, J. McCafferty, and S. Balasubramanian, *Quantitative visualization of DNA G-quadruplex structures in human cells*. *Nat Chem*, 2013. **5**(3): p. 182-6.
237. G. Biffi, M. Di Antonio, D. Tannahill, and S. Balasubramanian, *Visualization and selective chemical targeting of RNA G-quadruplex structures in the cytoplasm of human cells*. *Nat Chem*, 2014. **6**(1): p. 75-80.
238. G. Biffi, D. Tannahill, J. Miller, W.J. Howat, and S. Balasubramanian, *Elevated levels of G-quadruplex formation in human stomach and liver cancer tissues*. *PLoS One*, 2014. **9**(7): p. e102711.
239. R.A. Cardullo, S. Agrawal, C. Flores, P.C. Zamecnik, and D.E. Wolf, *Detection of nucleic acid hybridization by nonradiative fluorescence resonance energy transfer*. *Proceedings of the National Academy of Sciences of the United States of America*, 1988. **85**(23): p. 8790-8794.
240. G.A. Jones and D.S. Bradshaw, *Resonance Energy Transfer: From Fundamental Theory to Recent Applications*. *Frontiers in Physics*, 2019. **7**.
241. M.K. Johansson, H. Fidder, D. Dick, and R.M. Cook, *Intramolecular Dimers: A New Strategy to Fluorescence Quenching in Dual-Labeled Oligonucleotide Probes*. *Journal of the American Chemical Society*, 2002. **124**(24): p. 6950-6956.
242. S.A.E. Marras, F.R. Kramer, and S. Tyagi, *Efficiencies of fluorescence resonance energy transfer and contact-mediated quenching in oligonucleotide probes*. *Nucleic Acids Research*, 2002. **30**(21): p. e122-e122.
243. C.E. Rowland, J.B. Delehanty, C.L. Dwyer, and I.L. Medintz, *Growing applications for bioassembled Förster resonance energy transfer cascades*. *Materials Today*, 2017. **20**(3): p. 131-141.
244. A. De Cian, L. Guittat, M. Kaiser, B. Saccà, S. Amrane, A. Bourdoncle, P. Alberti, M.-P. Teulade-Fichou, L. Lacroix, and J.-L. Mergny, *Fluorescence-based melting assays for studying quadruplex ligands*. *Methods*, 2007. **42**(2): p. 183-195.

245. H. Tateishi-Karimata, K. Kawauchi, and N. Sugimoto, *Destabilization of DNA G-Quadruplexes by Chemical Environment Changes during Tumor Progression Facilitates Transcription*. *J Am Chem Soc*, 2018. **140**(2): p. 642-651.
246. M. Chen, Q. Chen, Y. Li, Z. Yang, E.W. Taylor, and L. Zhao, *A G-quadruplex nanoswitch in the SGK1 promoter regulates isoform expression by K(+)/Na(+) balance and resveratrol binding*. *Biochim Biophys Acta Gen Subj*, 2021. **1865**(2): p. 129778.
247. Y. Luo, A. Granzhan, D. Verga, and J.-L. Mergny, *FRET-MC: A fluorescence melting competition assay for studying G4 structures in vitro*. *Biopolymers*, 2021. **112**(4): p. e23415.
248. M. Dobrovolná, N. Bohálová, V. Peška, J. Wang, Y. Luo, M. Bartas, A. Volná, J.-L. Mergny, and V. Brázda, *The Newly Sequenced Genome of Pisum sativum Is Replete with Potential G-Quadruplex-Forming Sequences—Implications for Evolution and Biological Regulation*. *International Journal of Molecular Sciences*, 2022. **23**(15).
249. N.R. Markham and M. Zuker, *UNAFold*, in *Bioinformatics: Structure, Function and Applications*, J.M. Keith, Editor. 2008, Humana Press: Totowa, NJ. p. 3-31.
250. Y. Luo, D. Verga, and J.L. Mergny, *Iso-FRET: an isothermal competition assay to analyze quadruplex formation in vitro*. *Nucleic Acids Res*, 2022. **50**: p. e93.
251. L. Bonnat, L. Bar, B. Gennaro, H. Bonnet, O. Jarjayes, F. Thomas, J. Dejeu, E. Defrancq, and T. Lavergne, *Template-Mediated Stabilization of a DNA G-Quadruplex formed in the HIV-1 Promoter and Comparative Binding Studies*. *Chemistry*, 2017. **23**(23): p. 5602-5613.
252. L. Bonnat, M. Dautriche, T. Saidi, J. Revol-Cavalier, J. Dejeu, E. Defrancq, and T. Lavergne, *Scaffold stabilization of a G-triplex and study of its interactions with G-quadruplex targeting ligands*. *Org Biomol Chem*, 2019. **17**(38): p. 8726-8736.
253. E. Puig Lombardi, A. Holmes, D. Verga, M.P. Teulade-Fichou, A. Nicolas, and A. Londono-Vallejo, *Thermodynamically stable and genetically unstable G-quadruplexes are depleted in genomes across species*. *Nucleic Acids Res*, 2019. **47**(12): p. 6098-6113.
254. A. Guédin, A. De Cian, J. Gros, L. Lacroix, and J.-L. Mergny, *Sequence effects in single-base loops for quadruplexes*. *Biochimie*, 2008. **90**(5): p. 686-696.

Abstract in English

Guanine-rich nucleic acid sequences (both DNA and RNA) can generate four-stranded, noncanonical secondary structures, so called G-quadruplexes (G4). Generally, a G4 structure can differ by molecularity and be constituted by 1, 2 or 4 G-rich sequences. In this structure four guanine bases are connected by Hoogsteen hydrogen bonds (**Figure 17a**) to form a *G-tetrad*, or *G-quartet*, and two or more G-tetrads assemble to form a G4 structure. G-quadruplexes are polymorphic structures. Depending on strand orientation and the *anti* / *syn* conformations of guanines, G4 can adopt parallel, antiparallel and hybrid topologies (**Figure 17b**). G4 structures have been identified in genomes of different species and are involved in physiological processes, such as telomere maintenance, oncogene expression and genome instability. G4 may also participate in species evolution.

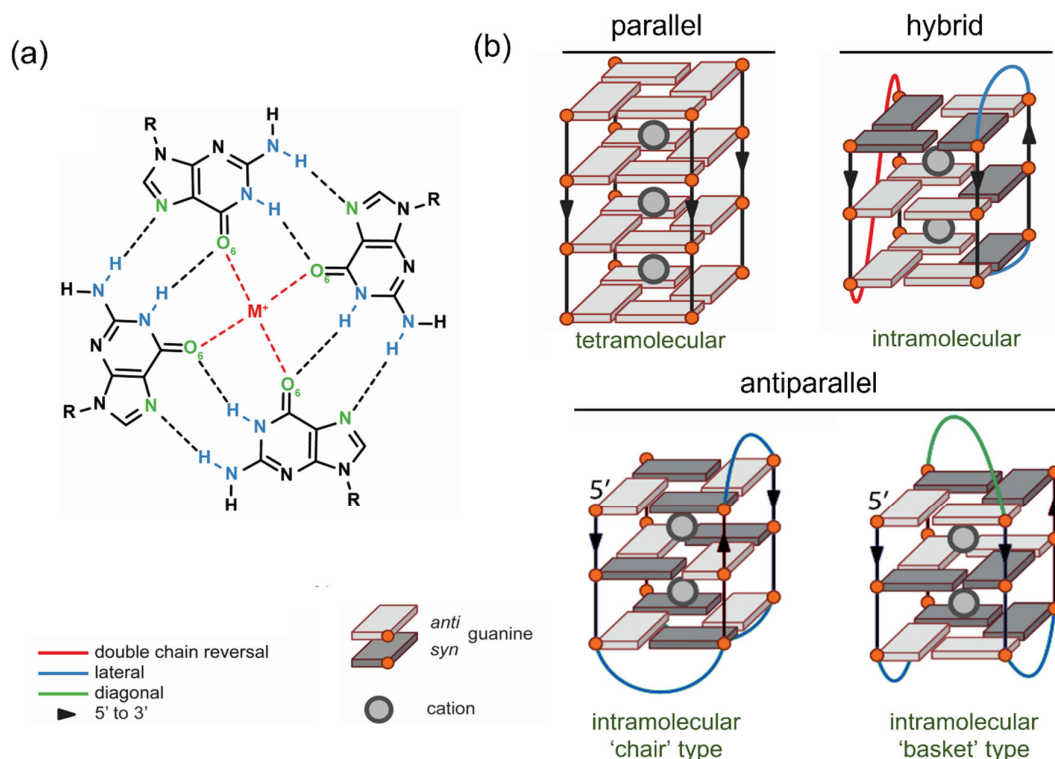


Figure 17 (a) Guanine tetrads are composed of four guanines that are linked together by eight Hoogsteen hydrogen bonds. (b) examples of G-quadruplexes topologies.

G4 prediction algorithms and G4 chromatin immunoprecipitation sequencing (G4 ChIP-seq) have been applied to mining numerous putative G-quadruplex sequences (PQS) in genomes. However, finding a sequence relatively rich in guanines does not necessarily mean it can form a G4 structure *in vitro*. Several well-established biophysical methods can be used to validate G4 structures *in vitro*. However, the majority of these techniques are either sample- or time-consuming, making them hardly suitable for characterizing a large number of sequences of PQS. The aim of this PhD thesis is mainly focus on the

development of novel *in vitro* high-throughput G4 characterization assays, suitable for validating a vast amount of PQS identified by *in silico* G4 prediction methods and CHIP-seq.

In the first work, we modified a classical fluorescence quenching-based melting assay to characterize intramolecular G4 structures. The two ends of a G4 sequence were labeled with a fluorophore and an appropriated quencher, respectively. At low temperature, the two labeled ends of an intramolecular G4 are located in close proximity and the fluorescence signal of the fluorophore is quenched by FRET. At higher temperatures, G4 unfolding occurs, causing the two ends to move far apart; thus, the FRET process does not occur and the fluorescence of the donor can be observed. The traditional FRET-melting assay can be performed in multiple-microwell plates by using a real time PCR instrument. However, in this assay, the reporter sequence should be dual-labeled, which is much more expensive than non-labeled counterparts. Additionally, not all the G4 unfolds (*i.e.*, parallel tetra-molecular G4s) change the distance of 5' and 3' termini.

Therefore, to characterize the structure of an unknown competitor we decided to tackle this issue by constructing a novel method, which was named the FRET-melting competition (FRET-MC) assay. A 21 mer telomeric intramolecular G4 was dual-labeled by the chromophore (FAM) and the quencher (TAMRA), abbreviated as F21T. PhenDC3 is a bisquinolinium derivative G4 ligand, which shows excellent affinity to all topologies of G4 structures and exhibits low affinity for non-G4 structures such as ssDNA, ssRNA, and dsDNA. We took advantage of the high selectivity of PhenDC3 towards G4 structures to challenge the interaction between F21T and PhenDC3 with unlabeled G4 competitors. PhenDC3 stabilizes the G4 structure of F21T and leads to an increase of F21T T_m of about 23 °C. An excess of G4 competitors can trap PhenDC3 and decrease the T_m of F21T back to the T_m value of F21T alone, while non-G4 competitors have little influence on the F21T-PhenDC3 interaction. FRET-MC has been validated with a reference panel of 65 known sequences. However, FRET-MC suffers from one non-negligible drawback, as it cannot be used to pin point G4s with low thermal stability, as these weak G4 behave as ssDNA at the temperature where F21T starts to melt.

To overcome this issue, an isothermal version of the competition was developed. Similar to FRET-MC, iso-FRET is based on the FRET mechanism and the interaction of PhenDC3 and a G4-forming quencher (37Q), while the fluorophore is bound to the short C-rich strand (F22) which is complementary to 37Q. When the competitor is a non-G4 sequence, PhenDC3 binds to 37Q and stabilizes its G4 formation, preventing the generation of the 37Q-F22 complex, leaving ON the fluorescence of F22. On the contrary, an excess of G4 competitor sequesters PhenDC3, letting 37Q and F22 hybridize leading to fluorescence quenching. The iso-FRET assay has been validated by 70 known sequences, including both DNA and RNA strands. Then the assay was used to investigate 23 different viral sequences, and the results were confirmed by other independent biophysical techniques (CD, IDS, TDS and FRET-MC). Of note, iso-FRET is not a perfect assay to characterize G4s *in vitro*: the excess of a G4-forming competitor characterized by high complementarity with F22 may form competitor-F22 duplex, preventing

fluorescence quenching and resulting in false negative results. We defined the *CF factor* to predict the complementarity level between the competitor and F22, and the results have shown that the majority of competitors can be well characterized by the iso-FRET assay; competitors with high risk to form a complex with F22 can be identified from the *CF*, calculated prior performing the experiment.

In parallel, a study focusing on characterizing the conformation of 'special' sequences which are rich both in cytosines and guanines was conducted. According to bioinformatics studies, there are over 11,000 PQS in the human genome (hg19) comprised at least two-three continuous cytosines, which may form a duplex instead of a G4 structure in a potassium deficient environment. A well-known human minisatellite G4 structure, CEB25 (PDB: 2LPW), is taken as a model sequence to study G4 formation in potassium / sodium mixed ion buffers, which constitute better models of physiological conditions than the universal-used mono cationic environments. Cytosines are used to gradually replace other bases in the long central loop of CEB25, to obtain a set of G4 mutated sequences. UV-melting results of the CEB mutated sequences showed that T_m values only change with the buffer rather than with the different cytosine-contained mutations, implying that cytosines in G4s are not preventing G4 folding, but rather increasing the possibility of duplex formation. CD spectra illustrated that changing potassium / sodium ratio could convert the parallel conformation of CEB sequences into other types of G4s.

Key Words: Nucleic acids structures, G-quadruplexes, FRET, Biophysical characterizations.

Resumé en français

Les séquences d'acides nucléiques (ADN et ARN) riches en guanine peuvent former des structures secondaires non canoniques appelées G-quadruplexes (G4). Généralement, ces structures G4s peuvent différer par leur molécularité en étant constituées de 1, 2 ou 4 brins riches en guanines. Dans cette structure, quatre guanines interagissent ensemble par des liaisons hydrogènes de type Hoogsteen pour former une G-tétrade (aussi appelée G-quartet), et deux ou plus G-tétrades s'empilent pour former la structure G4. Les G-quadruplexes sont des structures très polymorphiques, qui peuvent, en fonction de l'orientation des brins et la conformation *anti/syn* des guanines, adopter différentes topologies : parallèle, antiparallèle et hybride (**Figure 18**). Les structures G4s ont été identifiées dans le génome de différentes espèces et sont impliquées dans de nombreux processus physiologiques, comme la maintenance des télomères, l'expression d'oncogènes, et l'instabilité génomique. Les G4s peuvent aussi participer au processus évolutif.

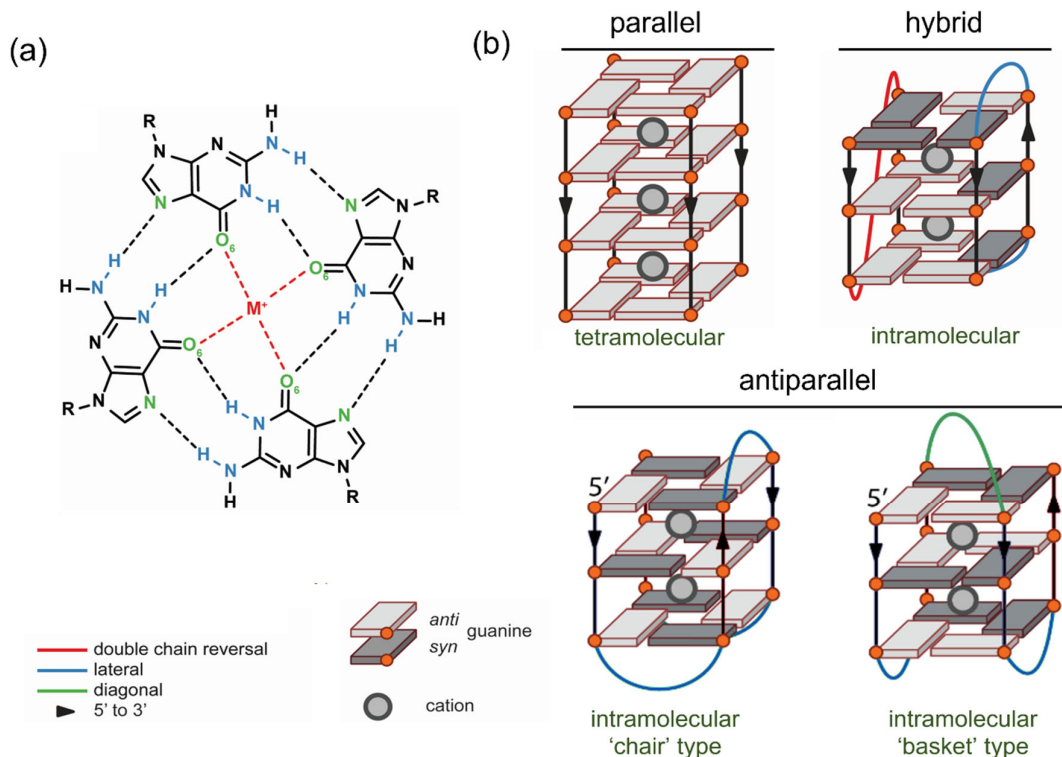


Figure 18 (a) Les tétrades sont composées de quatre guanines interagissent ensemble par huit liaisons hydrogènes d'Hoogsteen. (b) Exemples de topologies des G-quadruplexes.

L'identification de nombreuses séquences G-quadruplexes putatives (PQS) dans différents génomes a été possible grâce à des études bioinformatiques permettant de prédire les séquences G4s par différents algorithmes, mais également par séquençage de séquence G4s chromatinien isolé par immunoprécipitation (G4 ChIP-seq). Cependant, l'identification d'une séquence riche en guanines ne signifie pas forcément qu'elle peut former une structure G4 *in vitro*, et de nombreuses méthodes

biophysiques existent déjà pour les valider. Par contre, la majorité de ces méthodes sont consommatrices en temps et en échantillons, ce qui ne les rend pas pratique dans le cas de la validation d'un très grand nombre de PQS. L'objectif de la thèse se concentre principalement sur le développement de nouvelles méthodes *in vitro* à haut débit pour la caractérisation de G4, compatibles avec la validation d'un grand nombre de candidats identifiés préalablement *in silico* ou par ChIP-seq.

Pour commencer, nous avons modifié la méthode classique de FRET-melting qui permet de mesurer la température de dénaturation des G4s, basée sur le principe du FRET avec un couple de fluorophores donneur et accepteur, pour caractériser la formation d'une structure G4 intramoléculaire. Les deux extrémités d'une séquence G4 sont fonctionnalisées d'un côté avec un fluorophore donneur, et de l'autre, avec un désactivateur (accepteur) compatible. À basse température, les deux sondes sont à proximité et la fluorescence est désactivée par FRET. Au contraire, à haute température le G4 se dénature et les deux extrémités s'éloignent, le FRET ne se réalise plus et permet l'exaltation de la fluorescence. En général, le FRET-melting est réalisé en microplaque en utilisant une PCR quantitative. Cependant, pour cette expérience, toutes les séquences étudiées doivent être doublement fonctionnalisées, ce qui les rend plus coûteux à utiliser. De plus, la dénaturation de certains G4s intermoléculaires ne permet pas une variation significative de la distance entre le 5' et le 3' terminal pour être exploitable.

Par conséquent, nous avons décidé de développer une nouvelle méthode, appelé FRET-melting compétition (FRET-MC), pour la caractérisation de séquences inconnues par compétition. Pour cela, un G4 télomérique intramoléculaire de 21 nucléotides (F21T) a été doublement fonctionnalisé avec un fluorophore donneur (FAM) et un accepteur (TAMRA). Le PhenDC3 est un ligand G4 dérivé de la famille des bisquinolinium, montrent une excellente affinité pour toutes les topologies des G4s et presque aucune affinité pour les structures non G4s comme l'ADN simple brin, l'ADN double brin et l'ARN simple brin. Nous avons exploité cette sélectivité du PhenDC3 pour les structures G4 en réalisant la compétition de l'interaction entre F21T et le PhenDC3 contre un compétiteur non fonctionnalisé. Le PhenDC3 stabilise la structure G4 de F21T qui se traduit par l'augmentation de sa température de dénaturation d'environ 23°C. Un excès de compétiteur G4 piège le PhenDC3 et l'empêche de stabiliser F21T, ce qui provoque le retour du T_m à la température de dénaturation observée sans ligand. En revanche, un compétiteur non G4 n'aura aucune influence sur l'interaction et la stabilisation du PhenDC3 avec F21T. Le FRET-MC a été validé sur 65 séquences connues. L'un des inconvénients du FRET-MC est de ne pas être compatible avec des G4s moins stables thermiquement que F21T. En effet, ces séquences sont sous forme d'ADN simple brin avant même d'atteindre les températures où F21T commence à se dénaturer et donc aucune compétition n'est possible.

Pour répondre à cet inconvénient, une alternative isotherme a été développée. Similaire au FRET-MC, l'iso-FRET se base toujours sur le mécanisme du FRET, l'interaction du PhenDC3 avec une structure G4 fonctionnalisée avec un désactivateur (37Q), tandis que le fluorophore est lié à une séquence riche en cytosine (F22) partiellement complémentaire de 37Q. Si le compétiteur est une séquence ne formant

pas une structure G4, le PhenDC3 va se lier et stabiliser la forme G4 de 37Q, ce qui prévient l'hybridation de 37Q avec F22 et la fluorescence de F22 reste active. Au contraire, un excès de compétiteur G4 va séquestrer le PhenDC3, ce qui permet l'hybridation entre 37Q et F22 résultant en la désactivation de la fluorescence de F22. L'iso-FRET a été validé sur 70 séquences d'ADNs et d'ARNs connues. La méthode a également été utilisée pour étudier 23 séquences virales différentes, et les résultats ont pu être confirmés par d'autres techniques biophysiques indépendantes (CD, IDS, TDS et FRET-MC). Il est à noter que l'iso-FRET n'est pas parfait pour caractériser un G4 *in vitro* : un compétiteur G4 en excès peut possiblement s'hybrider avec F22 si la séquence est suffisamment complémentaire, ce qui résulterait en un faux négatif. Pour éviter cela, nous avons défini le facteur CF pour prédire le niveau de complémentarité entre le compétiteur et F22, dont les résultats montrent que la majorité des compétiteurs peuvent être caractérisés avec l'iso-FRET et les compétiteurs à risque peuvent être identifiés et exclus avant l'expérience par le calcul du CF.

Pour finir, nous nous sommes concentré sur l'étude de la caractérisation de séquences particulières riches en cytosines et guanines. Des études bioinformatiques ont montré que plus de 11.000 PQS dans le génome humain (hg19) contenaient au moins deux à trois cytosines contiguës, pouvant alors former un ADN duplex à la place d'une structure G4 dans un environnement déficient en potassium. Le minisatellite humain CEB25 (PDB: 2LPW) peut former une structure G4 et a été utilisé comme modèle pour étudier la formation des G4s dans un tampon mixte d'ions potassium et sodium, ce qui représente un tampon plus proche des conditions physiologiques plutôt que le tampon mono cationique normalement utilisé. Une série de séquences CEB25 mutées ont été développées où les cytosines remplacent progressivement les autres bases au niveau de la longue boucle centrale. Les résultats d'UV-melting avec les différentes séquences de CEB muté montrent que la valeur du T_m change uniquement en fonction du tampon et non pas avec le remplacement successif des bases de la boucle centrale par des cytosines. Cela implique que les cytosines dans la séquence ne déstabilisent pas la formation de la structure G4, mais augmentent plutôt la possibilité de former un duplex. De plus, l'étude par CD montre que le changement en proportion de potassium et de sodium dans le tampon peut conduire à un changement partiel de la conformation des séquences CEB de la forme parallèle vers d'autres types de conformations de G4.

Mots clés : Structure des acides nucléiques, G-quadruplexes, FRET, Caractérisation biophysique.