G4-Hunter : a new algorithm for G-quadruplexes prediction's

Amina Bedrat

université de BORDEAUX

THÈSE PRÉSENTÉE

POUR OBTENIR LE GRADE DE

DOCTEUR DE

L'UNIVERSITÉ DE BORDEAUX

ÉCOLE DOCTORALE : Sciences de la Vie et de la Santé

SPÉCIALITÉ : Bioinformatique

Par : Amina Bedrat

G4-Hunter : un nouvel algorithme pour la prédiction des G-quadruplexes

Sous la direction de : Dr. Jean-Louis Mergny

Soutenue le : 06/11/2015
Mention : Très honorable

Membres du jury :

Pr. Roger Marthan	- Université de Bordeaux (France)	Président
Dr. Geneviève Pratviel	- Université de Toulouse (France)	Rapporteur
Pr. Jean-Pierre Perreault	- Université de Sherbrooke (Canada)	Rapporteur
Dr. Patrizia Alberti	- CNRS/ INSERM/ MNHN (France)	Examinateur
Dr. Marie Beurton-Aimar	- Université de Bordeaux (France)	Examinateur
Dr. Jean-Christophe Andrau	- CNRS Montpellier (France)	Examinateur

G4-Hunter : un nouvel algorithme pour la prédiction des G-quadruplexes.

Des séquences compatibles avec la formation de G4 sont présentes au niveau de certaines régions clé du génome telles que les extrémités des chromosomes, mais également les régions de commutation de classe des immunoglobulines, les promoteurs de certains gènes dont des oncogènes et des séquences transcrites. Plus de 370 000 cibles potentielles ont été prédites lors des analyses bioinformatique du génome humain. Cependant, ces prédictions ne sont pas exhaustives étant limitées par la formulation des algorithmes de prédictions utilisés. En effet, les séquences recherchées suivent la formule consensus suivante $G_{3+}N_{(1-7)}G_{3+}N_{(1-7)}G_{3+}N_{(1-7)}G_{3+}$. Ainsi, en apportant plus de souplesse dans la description du quadruplex nous pourrons identifier et localiser plus de cibles potentielles. C'est pourquoi, nous proposons un nouvel algorithme G4-Hunter qui permettra l'identification la plus exhaustive possible de séquences cibles en prenant en compte la totalité de la région et non plus uniquement la cible potentielle. Par ailleurs, une étude expérimentale à grande échelle (sur une centaine de séquences cibles) a été menée afin de valider et tester la robustesse de G4-Hunter. A l'aide de ce nouvel outil, nous avons pu identifier de nouvelles séquences cibles non identifiées par les approches déjà existantes au sein des génomes humain, HIV et *Dictyostelium discoideum*.

Mots clés : [**Algorithme, Bioinformatique, G-quadruplex, ADN, Mitochondrie.**]

G4-Hunter: a new algorithm for G-quadruplexes prediction's.

Biologically relevant G4 DNA structures are formed throughout the genome including immunoglobulin switch regions, promoter sequences and telomeric repeats. They can arise when single-stranded G-rich DNA or RNA sequences are exposed during replication, transcription or recombination. Computational analysis using predictive algorithms suggests that the human genome contains approximately 370 000 potential G4-forming sequences. These predictions are generally limited to the standard $G_{3+}N_{(1-7)}G_{3+}N_{(1-7)}G_{3+}N_{(1-7)}G_{3+}$ description. However, many stable G4s defy this description and escape this consensus; this is the reason why broadening this description should allow the prediction of more G4 loci. We propose an objective score function, G4-hunter, which predicts G4 folding propensity from a linear nucleic acid sequence. The new method focus on guanines clusters and GC asymmetry, taking into account the whole genomic region rather than individual quadruplexes sequences. In parallel with this computational technique, a large scale *in vitro* experimental work has also been developed to validate the performance of our algorithm *in silico* on hundred of different sequences. G4-hunter exhibits unprecedented accuracy and sensitivity and leads us to reevaluate significantly the number of G4-prone sequences in the human genome. G4-hunter also allowed us to predict potential G4 sequences in HIV and *Dictyostelium discoideum*, which could not be identified by previous computational methods.

Keywords : [**Algorithm, Bioinformatic, G-quadruplex, DNA, Mitochondria.**]

Inserm U1212 (U869)

[Laboratoire ARNA, U1212, **2, rue Robert Escarpit 33607 Pessac, France**]

Resumé

Les acides nucléiques (ADN et ARN) sont des composants cellulaires essentiels au développement et au fonctionnement des organismes dont toutes atteintes (mutation, délétion, insertion ...) sont impliquées dans de nombreuses pathologies.
Depuis 1879, la composition des acides nucléiques est connue : sucre (désoxyribose et ribose respectivement pour ADN et ARN), phosphate et bases azotées (Adénosine A, Thymine T, Cytosine C, Guanine G), ce n'est qu'en 1953 que James Watson et Francis Crick déterminent la structure de l'ADN. Ils proposent le modèle de la double-hélice, qui est l'appariement de deux brins anti parallèle par des liaisons hydrogène entre les bases A-T et G-C.

Cependant, les acides nucléiques peuvent adopter des structures différentes. Elles sont dues aux différents arrangements possibles entre les bases azotées tels que l'appariement Watson-Crick inversé, l'appariement bancal et les liaisons Hoogsteen ou Hoogsteen inversé.
Parmi ces structures nous retrouvons:

- Simple-brin; structure tige-boucle ou épingle à cheveux,
- Double-brins; conformation A, B et Z de l'ADN,
- Triplexes; formés par l'insertion d'un simple-brin dans le sillon d'une double hélice et constitués d'appariement de type Watson-Crick et Hoogsteen,
- Quadruplexes; repliement d'un, deux ou quatre brins d'ADN ou ARN.

Ainsi, les séquences ADN ou ARN riches en guanines peuvent adopter une structure différente de la double-hélice classique, et se replier en structures appelées quadruplexes ("G4"). Cette conformation repose sur la formation de quartets de guanines, où quatre guanines coplanaires établissent un réseau cyclique de liaisons hydrogène. Il a été démontré que la stabilité de ces structures est assurée par la nature et la concentration du cation (K^+, Na^+) présent entre les quartets consécutifs.
Par ailleurs, ces structures présentent un polymorphisme structurel important dû entre autre à la molécularité et l'orientation des brins.

Les G-quadruplexes peuvent être constitués de quatre (intermoléculaires), de deux (dimer) ou d'un seul brin (mono- ou intramoléculaire) d'ADN.
L'orientation des brins est également responsable de la diversité structurale des G4. Quatre possibilités sont décrites; (i) les quatre brins sont orientés dans la même direction (G4 parallèle), (ii) un des brins est orienté dans la direction opposée des autres (G4 "hybride" ou (3+1)),

(iii) les deux brins voisins ou diamétralement opposés sont orientés dans une direction et les 2 autres dans la direction opposée (G4 anti parallèle).

Des séquences compatibles avec la formation de G4 sont présentes au niveau des régions répétées des télomères, trouvées aux extrémités des chromosomes, et le promoteur de nombreux oncogènes tel que : c-myc, c-kit [51][50]. Il est supposé qu'elles jouent un rôle important dans la régulation de la transcription, la translation[222] et la régulation de l'expression des gènes.
Les G4 ne sont pas uniquement spécifiques aux eucaryotes. Ainsi, des séquences avec le potentiel de former des G4 ont été identifiées au sein des virus. En effet, le génome du papillomavirus (HPVs) présente sept séquences avec un potentiel de former un G4 stable.

Les G-quadruplexes semblent jouer ainsi un rôle dans de nombreux processus cellulaires tels que le contrôle de l'expression génique, la réplication ou l'épissage et représenteraient ainsi des cibles thérapeutiques d'intérêt. D'où la nécessité d'avoir à disposition des outils fiables permettant d'identifier les séquences à fort potentiel *in silico*.

Historiquement, le premier algorithme de prédiction développé est Quadparser [19].
Quadparser[1] repose sur l'identification du motif prédéfini, d($G_{3+}N_{(1-7)}G_{3+}N_{(1-7)}G_{3+}N_{(1-7)}G_{3+}$) (N= A, T, C ou G). Ainsi les séquences identifiées sont des G4 intramoléculaires, avec trois quartets et plus, sans discontinuité dans les blocs de guanine et des boucles de 7 bases. Les séquences prédites peuvent potentiellement former un G4 en conditions physiologique (100mM KCl, 10mM Tris-HCl (pH 7.4))
Ecrit en language C, Quadparser identifie de nouvelles séquences à partir de données génomiques au format FASTA. Le format de sortie et les motifs recherchés peuvent être personnalisés.
Ainsi, plus de 370000 séquences ont été identifiées lors de l'analyse du génome humain.

D'autres algorithmes accessible en ligne peuvent également être utilisés. QGRS Mapper (Quadruplex forming G-Rich Sequences) est l'un de ces algorithmes [92]. L'interface d'analyse permet à l'utilisateur de moduler le motif recherché en intervenant sur le nombre de guanines constituant les quartets (au minimum 2 selon les règles de l'algorithme), la taille de la séquence recherchée et non seulement la taille des boucles (0 à 36 par défaut) mais aussi la composition de celle-ci et cela par de simple expression régulière. L'algorithme attribue à chacune des séquences identifiées un score. Le calcule du score est dépendant de trois principes :

- Les petites boucles sont plus communes que les longues,
- Les G4 ont tendance à avoir des boucles de même tailles,
- Plus le nombre de bloc de guanine est élevé plus le G4 est stable.

Ainsi, plus le score est élevé plus la séquence prédite a un potentiel G4 élevé.
QGRS Mapper, écrit en PHP et JAVA (pour les graphiques), analyse des sequences d'ADN et d'ARN en utilisant les données de NCBI ou des séquences fournis par l'utilisateur au format brute ou FASTA. Les pages de sortie regroupent un ensemble d'information sous forme de tableaux et de graphiques en précisant la localisation des exons ainsi que la localisations des sequences identifiées.

[1] Disponible sur demande

L'interface web Quadfinder[2], permet non seulement de prédire de nouveaux G4 à l'aide de Quadparser mais également de définir la stabilité thermodynamique grâce au Bayesian learning algorithme. Contrairement aux algorithmes développés jusqu'ici, ddiQFP (duplex-derived interstrand Quadruplex Forming Potentiel) implémenté en Perl, permet d'identifier des G4 intermoléculaires présents au sein de sites ciblés par l'hélicase pif1 [100].
Beaudoin *et al* proposent un algorithme dédié à la recherche de G4 au sein des ARN [97]. Le programme calcule le score d'une séquence ARN en fonction du score des blocs de guanine/-cytosine. Cet algorithme est caractérisé par une grande sensibilité et spécificité et comme un outil complémentaire au outils déjà existants.

Ces études bioinformatiques ont identifié plusieurs séquences et démontré comment les quadruplexes sont distribués dans le génome. La diversité structurelle des quadruplexes observés, propose un motif différent de celui recherché par ces algorithmes ainsi le nombre de quadruplexes prédit peut être plus élevé.

En effet, ces études se basent sur des informations structurelles (recherche de motif) et des analyses comparatives de la localisation des ces séquences [84]. L'analyse est réalisée uniquement sur de petites séquences ce qui limite la détection de séquences avec de longues boucles qui peuvent former des G4 [16]. La séquence c-myc retrouvée au sein du promoteur de ce même gène présente un polymorphisme. En effet, dans certaines condition salines c-myc ne suit pas le motif rechercher [20]. De même des séquences avec plus de trois quartets peuvent également exister [101].
Les méthodes de calcule de score quant à elles se basent sur des caractéristiques de la séquence tels que la longueur de la boucle le nombre de groupe de guanines ou les séquences avoisinantes.

Clairement ces outils prédisent de nouvelles séquences en fonction de caractéristiques de séquences déjà existantes mais leurs prédictions sont limitées et un grand nombre de faux négatifs sont générés. La recherche d'un nouvel outils de prédiction et donc une nécéssaire.
Nous proposons un nouvel outil qui,

- Favorise les blocs de guanines,
- Favorise l'asymétrie G/C (recherche de G4 dans les deux brins),
- Repousse les limites des outils existent (longueur de la boucle, nombre de quartets),
- Recherche de motifs insolites.

A partir d'une idée simple, l'algorithme recherche des séquences formant des G4 par calcule de score, ce score dépende uniquement de la composition en bases de la séquence. En effet les bases de la séquence sont converties en chiffres (de -4 à +4 selon le nombre de guanines consécutives) et le score représente la moyenne d'une fenêtre de 25 à 100 nucléotides. Plus la valeur absolue du score est élevé plus la séquence est susceptible de former un G4.

[2]Indisponible en ligne

J'ai tout d'abord implémenté l'algorithme en Python prend en entrée un fichier au format FASTA (qui peut comprendre une ou plusieurs séquence FASTA) et produit deux fichiers texte (un fichier qui représente les scores de chaque fenêtres et un fichier qui regroupe les séquences chevauchantes avec leurs nouveau score) et une représentation graphique des scores obtenus.

j'ai ensuite analysé un ensemble de séquences formant ou non un G4, issue de la littérature et validés expérimentalement, afin de valider notre algorithme. Le score, pour une fenêtre de 25 nucléotides, est calculé pour ces deux groupes de séquences (G4, non G4) et pour une (P *value* <0.001) nous retrouvons une différence significative entre les deux distributions du score.

J'ai également analysé l'ADN mitochondrial avec l'algorithme que nous avons développé et comparé les séquences ainsi obtenus avec celles identifiées par Quadparser. Cent soixante sept candidats ont été identifié (avec un score > 1 et une fenêtre d'analyse de 25 nucléotides). En revanche, lors de l'analyse avec Quadparser (G= 2 à 5 et N = 1 à 7) 81 candidats ont été identifiés dont 23 uniquement identifié par cet algorithme. Ces séquences de plus petite taille que notre fenêtre d'analyse (< 25 nucléotides), ont un score inférieur à 1.

Une fois prédite il est nécessaire de valider expérimentalement la capacité des séquences à former un G4. De nombreuses techniques sont décrites. Au sein de notre équipe, nous utilisons principalement (i) les techniques de spectroscopie UV: dénaturation thermique (Thermal merlting-Tm) [103], spectres de différence thermiques (Thermal Difference Spectra TDS) [102], spectre de différences isotherme (Isothermal Difference Sperctra IDS) [109] et Dichroïsme Circulaire (CD) [107], (ii) Résonance Magnétique Nucléaire (RMN) [22] (iii) et un test de fluorescence développé au sein de l'équipe, le test de la thioflavine T [104].

J'ai réalisé une étude expérimentale de grande ampleur qui consiste à tester expérimentalement tous les nouveaux candidats. Ainsi 75% des séquences prédites forment un G4 pour score >1.
24% des candidats dont le score est entre 1 et 1.25 ne forme pas de G4 ou de conformation inconnue (nos tests biophysique n'ont pas été concluants) . Au delà d'un score de 1.5 les séquences ont tendance à former un G4 stable et au delà d'un score de 2 toutes les séquences forment des G4 stables.
Le nombre de G4 formés au sein de la mitochondrie est plus grand que le nombre de G4 prédit par Quadparser. Ainsi, le nouvel algorithme répond aux conditions déjà posées et propose une nouvelle liste de G4 testée et validée expérimentalement. A partir de cette étude, nous avons choisi la valeur 1 comme valeur seuil du score pour les nouvelles recherches.

En collaboration avec Amrane S., nous avons identifié 10 candidats suite à l'analyse d'un alignement de 1684 génomes d'HIV-1. La région conservée riche en guanine du promoteur du HIV-1, connue pour réguler sa transcription, forme un G4 antiparallèle stable [189]. Nous avons aussi breveté cette séquences pour son potentiel d'inhiber le virus d'HIV.

Enfin nous avons développé un nouvel outil bioinformatique dédié à la recherche de nouveaux G4, sa fiabilité est démontré par une analyse des deux génomes : la mitochondrie et HIV-1. Ainsi, Nous souhaitons pour la suite de cette thèse proposer une interface graphique qui rend l'utilisation de notre algorithme facile, améliorer les paramètres de recherche (fenêtre et seuil du score), rechercher de nouveaux G4 dans d'autre génomes tels que *Dictyostelium sp., Caenorhabditis elegans*, Ebola et Marburg.

Acknowledgements

I would like to thank the entirely and especially Merciful.
I thank my parents and my brother, the true love.
I thank my PhD supervisor, the big boss ever.
I thank Marie-Noël B.A. & Patricia T., the best souls that a foreigner can meet.
I thank my traveling friend, my lab brother and ma cherie.
I thank every single person I meet in my life.
I thank all the members of my thesis committee.

Contents

List of Figures

List of Tables

State of the art

Chapter 1

Guanine Quadruplexes

Contents

The story of DNA structure is as varied as it is interesting. The beauty and the simplicity of the DNA double-helix structure are all that is needed for a basic comprehension of cellular genetics. This double-helix was proposed by J.D. Watson & F.H.C. Crick in 1953 [1], based on the rules of Chargaff and the diffraction images obtained from R.E. Franklin [2]. DNA is usually represented by a right-handed double-helix formed with two anti-parallel strands held together by complementary base pairing (Fig.1.1-A). This pairing is made through Hydrogen bonds between the donor and acceptor "Watson-Crick" nucleotides sites of the bases A-T and C-G (Fig.1.1-B). The succession of base pairs (Fig.1.1-C & -D) defines the genetic information needed by the cells to accomplish their vital functions. The helix is stabilized by important staking interactions between consecutive bases. B-DNA has always been regarded as the biologically relevant structure; however DNA can adopt a wide variety of conformations including highly distorted A-form (observed under conditions of low hydration), Z-form (favoured at high salts concentration) and parallel-stranded DNA (Fig.1.1-A).

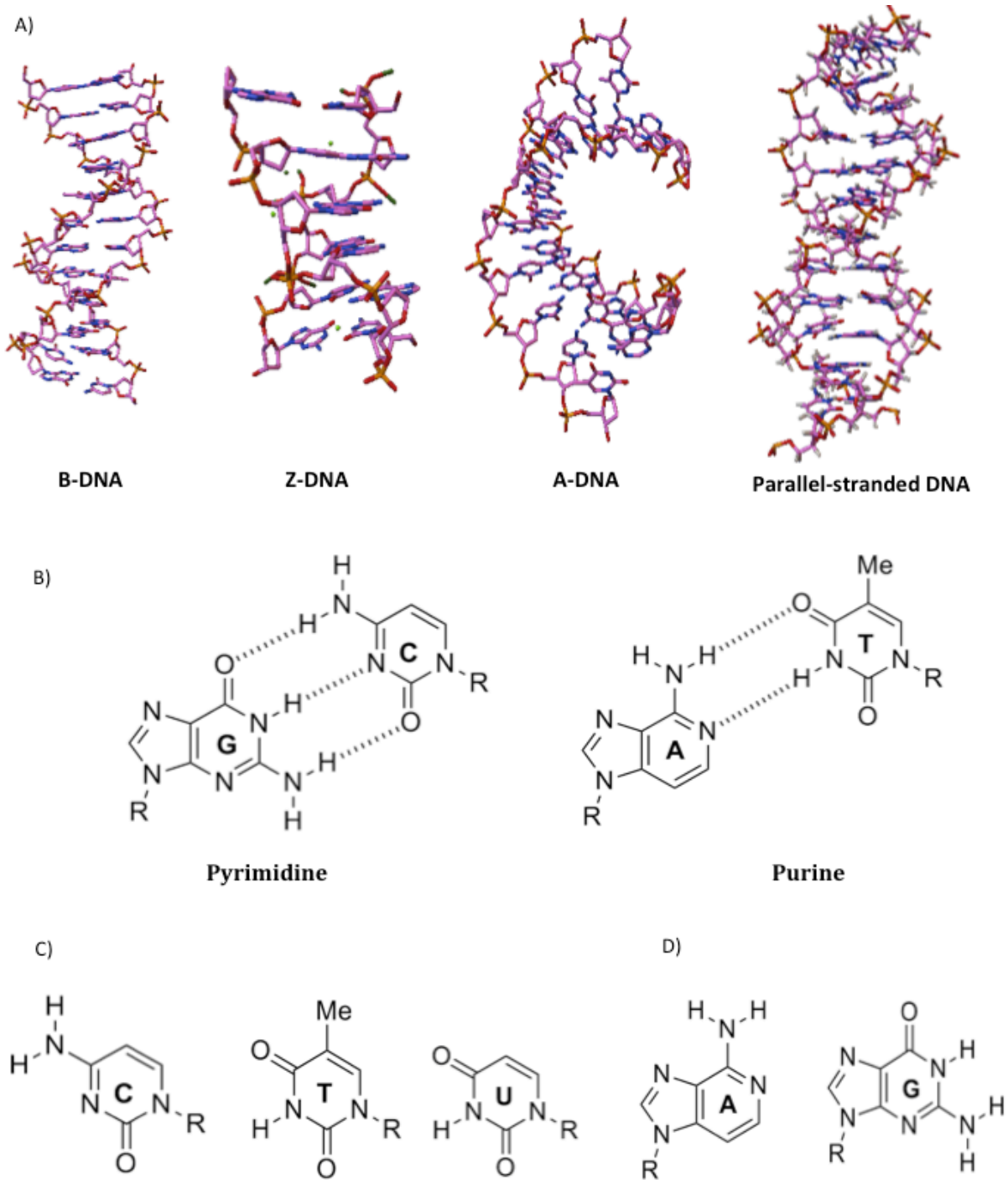


Figure 1.1: **Nucleic acids components & different duplex DNA structures.** (**A**) Models of the A, B and Z DNA. B and A forms are right-handed helices, B form (PDB : 1BNA) is characterized by a helical turn every 10 base pairs (3.4 nm) with 0.34 nm by base pair. A form (PDB : 440D) is more compact with 11 base pairs per turn. Z form (PDB : 4FS5) is left-handed with a zig-zag like backbone. Parallel-stranded duplex DNA (PDB : 1JUU) are maintained by reverse Watson-Crick base pairing between A-T or alternating, symmetrical self-pairs of G_{syn}-G_{syn} (**B**) Canonical base pairing (Watson-Crick) of nucleotides within the double-helix of DNA (A-T & C-G). (**C & D**) Schematic structures of the nucleobases constituting DNA and RNA.

Since Watson and Crick study, it is now well established that DNA can adopt different secondary structures. DNA can folds into a vast variety of hairpin, triplex, i-motif and G-Quadruplex structures containing non-canonical base pairs, triads or quartets.
Some of these non-canonical base pairing schemes are called "Hoogsteen" as they were highlighted by Karst Hoogsteen using X-ray diffraction [3]. They can be established within guanine-rich DNA sequences between guanines via four Hydrogen bonds involving the "Watson Crick" and "Hoogsteen" edges (Fig.1.2-A & -B).

Triplex DNA consists of a double-stranded DNA (dsDNA) (with one purine-rich and one pyrimidine-rich strand) and a single-stranded triplex-forming oligonucleotide (TFO) that binds to the major groove of the duplex through Hoogsteen or reverse Hoogsteen bonding with the purine-rich strand [4, 5] (Fig.1.2). Cytosine-rich sequences may adopt an i-motif structure (Fig.1.3) at acidic pH, which consists of two parallel-stranded DNA duplexes held together in an antiparallel orientation by intercalated, hemiprotonated cytosine–cytosine+ base pairs [6].

The G-quadruplexes are formed by guanine-rich sequences. This structure is based on the stacking of several tetrads, which were first identified in 1962 as the basis for the aggregation of 5'-guanosines monophosphate (GMP) [7]. Today, the interest for G4 is growing with thousands of reports on structure, detection within cells and function [8]. For many years G-quadruplexes were considered as a structural curiosity, but now come up in various areas, ranging from biology and medical biology to supramolecular chemistry and nanotechnology. In this chapter, I will describe the main characteristics of the non-canonical DNA "G-quadruplex" structures.

1.1 Folding and Topology of G-quadruplexes

G-quadruplexes are four stranded structures formed by guanine-rich DNA (or RNA) sequences. The association of four guanines by a network of eight Hydrogen bonds generates a planar structure called a *G-quartet* or G-tetrad (Fig.1.4-A) and the stacking of several G-quartets forms a G-quadruplex (Fig.1.4-B & -C) [7, 9]. The primary building block of this structure, the G-quartet, is composed of four coplanar guanines that interact with each other via Hoogsteen base pairs. The G-quartets are maintained by the presence of monovalent cations, mainly the metallic cations such as K^+, and to a lesser degree Na^+ [10], and others such as: Rb^+, Sr^{2+}, Ca^{2+} and Pb^{2+}[11]. The most common mode of cation binding is the "sandwich" mode where a cation is positionned between two G-quartets and coordinated to eight carbonyl O6 atoms, thereby minimizing their repulsion and providing stability to the quartet [12].

G-quadruplexes are highly polymorphic depending on the nucleic acid sequence and experimental conditions. The four strands serving as columns supporting the G-tetrad core can be in different orientation, G4 can be formed in an intermolecular or intramolecular fashion (Fig.1.5-Top). The loops connecting these strands (Fig.1.5-Down) can also adopt different conformations [13].

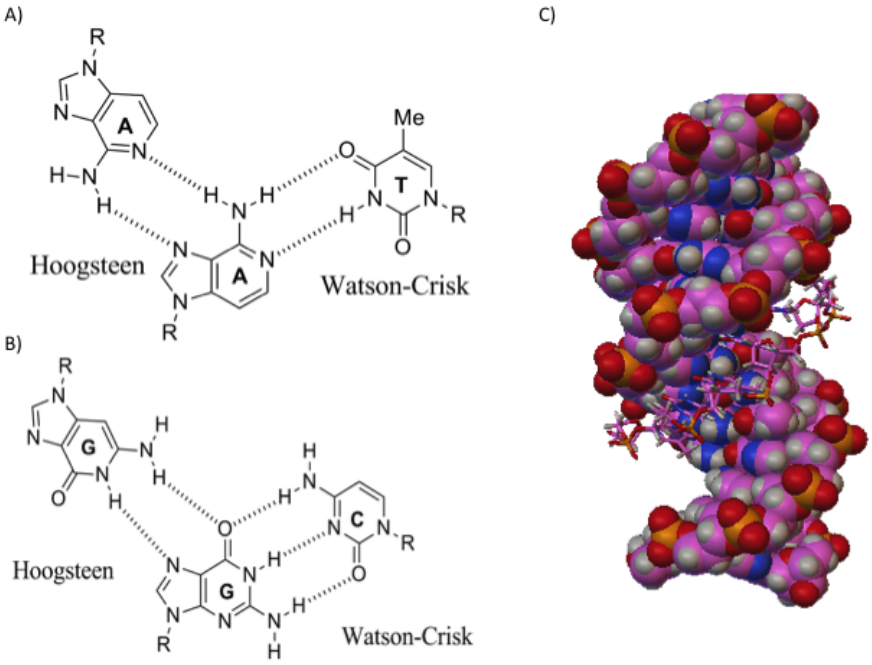


Figure 1.2: **Triplex DNA structure.** Schematic representation of (**A & B**) G-GC and A-AT triplets (PDB : 1BWG), involved in purine and pyrimidine triplexes, respectively. (**B**) Triplex DNA adapted from the structure (PDB : 1BWG). The triplex above is called a purine triplex as the TFO is purine rich. Other triplex may be formed with pyrimidine oligonucleotides, based on the formation of C^+-GC and T-AT triplets .

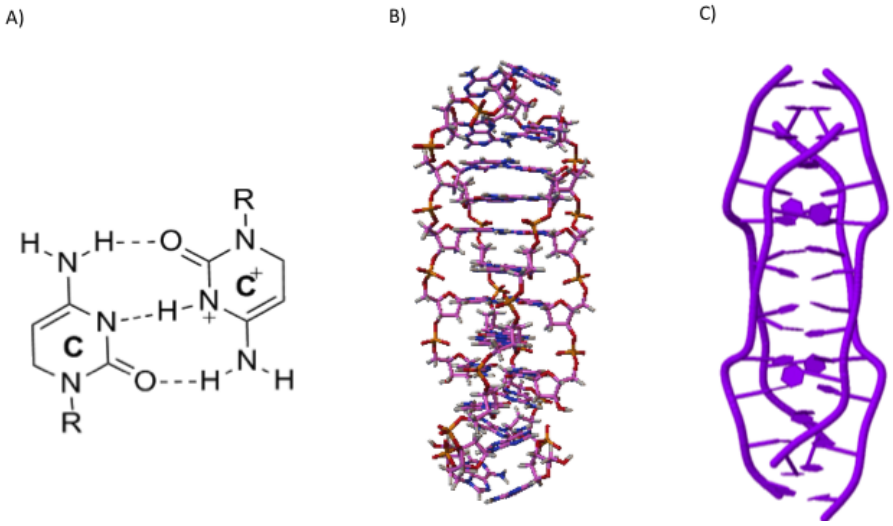


Figure 1.3: **I-motif DNA structure.** (**A**) A hemiprotonated cytosine–cytosine+ base pair. (**B & C**) Structure of the d(A2C4) intermolecular i-motif (PDB : 1YBL).

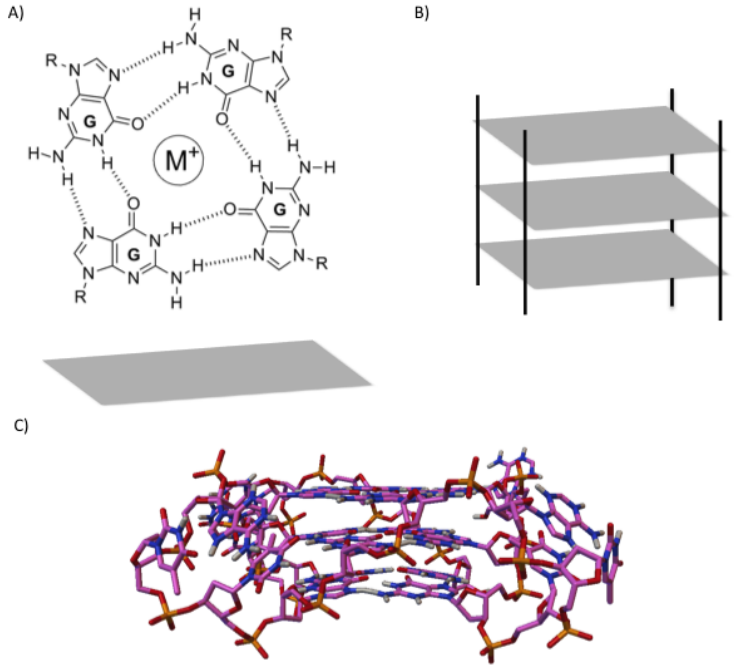


Figure 1.4: **G-quadruplex DNA structure.**(**A**) Presentation of a G-quartet formed by the association of four guanines. (**B**) G-quadruplex nucleic acids structure. (**C**) The intramolecular model of G4 formed by the stacking of three G-quartets adapted from the human telomeric DNA sequence (PDB : 1KF1).

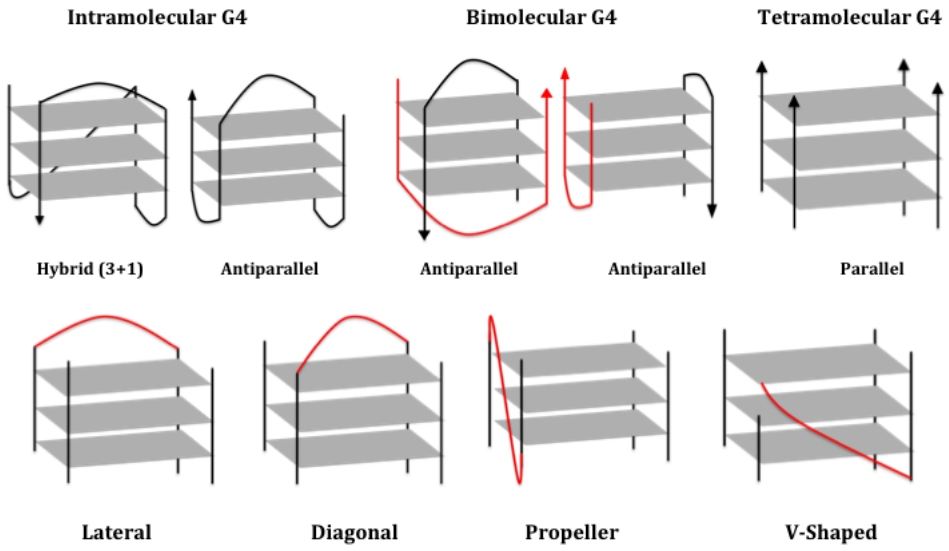


Figure 1.5: **G-quadruplex polymorphism.** Schematic representation of different G4 conformations according to the molecularity (intra-, bi-and tetra-molecular), strand orientations (parallel, anti-parallel and hybrid (3+1)) (top). Schematic representation of different loop conformations within G4 (lateral, diagonal, propeller and V shaped) (down).

1.2 G-quadruplex structural polymorphism

The stacked G-quartets linked by the phosphate backbone constitute the relatively invariant core of all G4 structures. Different parameters have been described to explain the structural diversity of G4 such as the number of strands, their orientations, the conformation of loops and the nature of cations.

1.2.1 Strands number & orientations

G-quartets can be assembled in an intramolecular (monomolecular) fashion, where one strand containing several blocks of guanines is able to fold back and form a G4 structure. G4 can also be formed from two-, three- or four-strands, resulting in intermolecular structures that can adopt a wide variety of conformations (Fig.1.5) [8].

Within each strand, the glycosidic conformations of guanines can be either *syn* or *anti*. The relative orientations of strands are geometrically related with the glycosidic conformation of guanines [8]. The possibilities are: (i) Parallel G4 (four strands identically oriented) (ii) "3+1" Hybrid G4: three strands oriented at the same direction and in the opposite direction, (iii) "2+2" Anti-parallel G4: two strands in one direction, the others in the reverse orientation. These antiparallel G4 can be subdivided into two distinct classes: two parallel strands can be adjacent or diagonal, and this will result in very different structures and groove sizes [14]. These backbone strands are connected by linkers commonly called loops.

1.2.2 Loop conformations

Monomolecular and bimolecular structures generally present 2 or 3 loops. The loops are linkers connecting G-stretches that support the G-tetrad core. Four conformations are distinguishable: (i) lateral or edgewise loops connecting two anti-parallel adjacent strands, (ii) diagonal loops connecting two opposing antiparallel strands, (iii) "Propeller" or Double-chain reversal loops (N-shape) connecting two adjacent parallel strands, and (iv) V-shaped or snap-back loop connecting two corners of G-tetrad core in which a support column is missing (the wedges of G-quartet).

Furthermore, loop residues can form base-pairing alignments, which in turn stack with the terminal G-tetrads, further stabilizing G-quadruplex structures [14]. The loop conformations is closely linked to strand orientations of a G4 and depends on the size and the sequence of the linkers. Different studies pointed the effect of loops length and composition on the G4 formation and stability. They indicate that: (i) loops consisting of a single nucleotide are generally N-shaped and promotes parallel stable G4 formation; (ii) loops of ≥ 3 nucleotides (nt) promotes an antiparallel conformation with a decrease in stability [15] and (iii) a sequences with a loop of more than 9 nt may still form a G4 [16, 17].

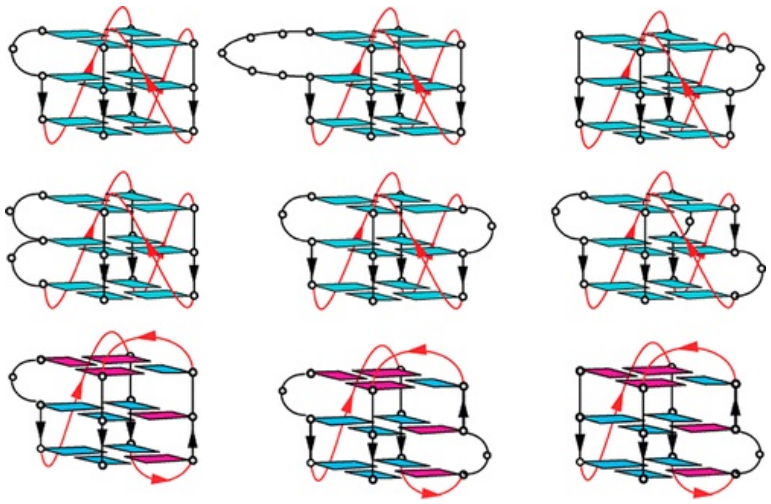


Figure 1.6: **Bulges in G-Quadruplexes: broadening the definition of G-Quadruplex-forming Sequences.** The bulges can be formed in many different situations within G-quadruplexes, thus making some G4 sequences defying the standard description. Adapted from [13].

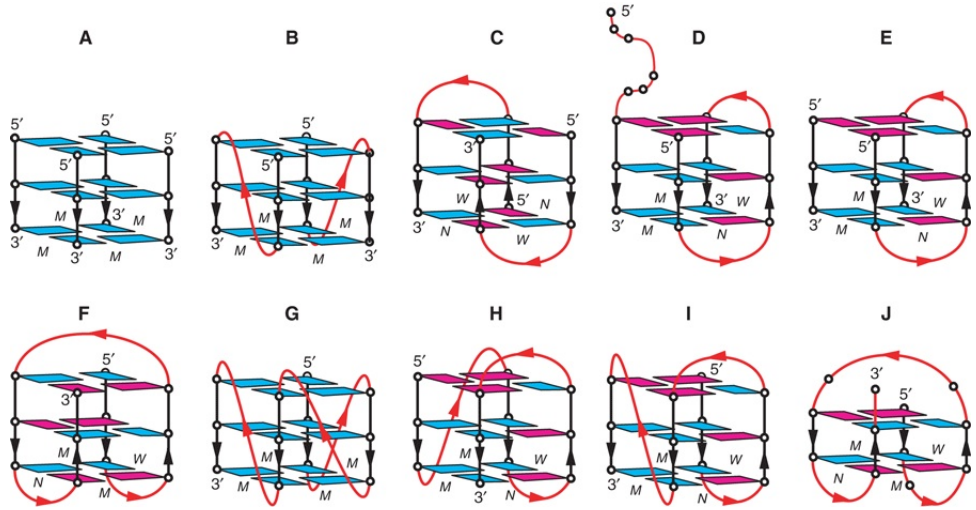


Figure 1.7: **Schematic structure of human telomeric G-quadruplexes.** (A) Tetrameric parallel-stranded G4 observed for the single-repeat human telomeric sequences d(TTAGGG) and d(TTAGGGT) in K^+ solution. (B) Dimeric parallel-stranded G4 observed for the two-repeat human telomeric sequence d(TAGGGTTAGGGT) in a K^+ containing crystal and in K^+ solution. (C) Dimeric antiparallel-stranded G4 observed for two-repeat human telomeric sequence d(TAGGGTTAGGGT) in K^+ solution. (D) Asymmetric dimeric (3 + 1) G4 observed for the three-repeat human telomeric sequence d(GGGTTAGGGTTAGGGT) in Na^+ solution. (E) Asymmetric dimeric (3 + 1) G4 association observed for the three-repeat human telomeric sequence d(GGGTTAGGGTTAGGGT) and the single-repeat human telomeric sequence d(TAGGGT) in Na^+ solution and in K^+ solution (unpublished results). (F) Basket-type form observed for d[A(GGGTTA)3GGG] in Na+ solution. (G) Propeller-type form observed for d[A(GGGTTA)3GGG] in a K^+ containing crystal. (H) (3 + 1) Form 1 observed for d[TA(GGGTTA)3GGG] in K^+ solution. (I) (3 + 1) Form 2 observed for d[TA(GGGTTA)3GGGTT] in K^+ solution. (J) Basket-type form observed for d[(GGGTTA)3GGGT] in K^+ solution. Adapted from [18].

1.2.3 G-quadruplex and bulges

The knowledge accumulated on G4 structures led to the formulation of a consensus motif, which is capable of forming an intramolecular G4 used in the search for Putative Quadruplex Sequences (PQS).

To date, a G-quadruplex is often considered to be formed by a sequence containing four tracts of two, three or more continuous guanines connected by linkers, in which the G-tracts would form continuous columns supporting the G-tetrad core, while the linkers would form loops connecting the corners of the G-tetrad core [19]. This motif is widely used and most of G4 structures resolved are conform to this consensus. Nevertheless, it has been shown that several examples defy the description of this consensus, such as the sequence Pu24 from the human c-myc promoter [20] and c-kit87up from the human c-kit promoter [21]. These sequences exhibit discontinuous arrangement of guanines in one column of the G-tetrad core, despite the presence of four G-tracts each having at least two or three continuous guanines [20, 21]. While a loop connects two corners of the G-tetrad core, bulges are projections of bases from the G-tetrad core: they connect two non-adjacent guanines of the same strand within the G-tetrad core.

Recently, Phan *et al* [13] have shown that many different bulges can exist in G4 structures. They vary in their sequence, size, position and numbers within the G4. This result alters the common view on the ability of many sequences to form G-quadruplexes. Expanding this description should help to identify more potential G4 forming sequences. It could also open the possibilities of exploiting bulges as recognitions elements for interaction between G-quadruplexes and other molecules. In the same view NMR-based solution structures have been reported for several sequences in which G-quartets are connected by *four* loops. The guanines involved in the G-tetrads include some isolated guanines but can exclude guanines from the G-tract (e.g. c-myc23456 and c-kit1) [22, 20].

1.2.4 RNA quadruplexes

Coding and non-coding RNA containing several blocks of guanines are able to fold into G-quadruplexes. It has also been shown that many G4 RNA possess remarkable stability [23, 24]. In addition, the formation of G4 RNA seems all more relevant since RNA is generally single-stranded. Such structures might play key roles in the gene regulation, translation, genomic stability and disease. The human fragile X syndrome is caused by: (i) the loss of FMRP protein the ability to bind RNA, or (ii) the repression of the gene FMR1 coding this protein. The *in vitro* selection showed that the FMRP protein bind to G4-prone sequences [25]. Bioinformatics studies demonstrated that the 5' and 3' untranslated regions (UTRs) of mRNA are G-rich. Guanine-rich UTRs can fold into G4, which may play a role in the regulation of translation, mutation and degradation of RNA [26, 19]. Recent studies reported that telomeric repeat-containing RNA (TERRA) folds into a parallel G4 conformation; additionally, the parallel RNA G4 tends to associate and form higher-order structures [27]. TERRA could play important roles in regulating telomerase and in chromatin remodeling during cell development [28].

1.3 Biological functions of G-quadruplexes

1.3.1 Localization of G-quadruplexes

Both prokaryotic [29] and eukaryotic genomes from yeast to human are rich in potential G4-forming motifs [30]. Among these are repetitive and functionally essential chromosomal domains, including telomeres, rDNA and the immunoglobulin heavy chain switch regions of higher vertebrates. Many minisatellite and microsatellite repeats are G-rich and have the potential to form G4 DNA. G-rich regions are also found within specific single-copy genes.

Telomeres

Telomeres are specialized nucleoprotein complexes that cap and protect the extremities of linear eukaryotic chromosomes from degradation [31]. During each cell division, telomeric DNA shortens by 50-200 base pairs, until a critical size is reached, causing cell division arrest. In most eukaryotes, telomeric DNA consists of a tandem array of a short motif of 5–8 nts, that includes two, three or four consecutive and highly conserved guanines (Fig 1.7)[32]. The GGGTTA motif is found in many phylogenetically distant organisms, including vertebrates [33], several fungi [34] and slime molds [35]. Variants of this motif are found in many other organisms: parasites [36], plants [37], algae [38], nematodes [39] and yeasts [40]. Such a structure might be important for telomere biology and a good target for drug design [14]. Since then, numerous structural studies were able to demonstrate the complexity of human telomeric sequence folding and conclude that this motif can be folded into parallel, hybrid or antiparallel G-quadruplexes depending on experimental conditions and exact sequence [18].

Telomeric repeats are normally capped by a protein complex that identifies them as telomeres rather than damaged DNA and protect them from misguided cellular efforts at repair that are potentially destabilizing [41]. The removal of this telomere-capping complex results in telomere instability that can be countered by drugs that stabilize G4 structures [42]. The shortening process is countered by telomerase [43], an enzyme required for the proliferation of stem cells and germline cells, as well as most cancer cells [44]. Telomerase is highly over-expressed in many cancer cell types whereas it is only expressed at low levels in normal somatic cells, allowing cancerous cells to be replicated indefinitely. Folding of telomeric DNA into G4 seems to influence the extent of telomere elongation *in vitro* and might therefore act as a negative regulator of elongation *in vivo* [45]. Therefore, G-quadruplexes formed by human telomeric DNA have been considered as promising anticancer targets [46].

Promoters & UTRs

Guanine-rich tracts are observed in critical segments of eukaryotic and prokaryotic genomes. Genome-wide computational analysis has revealed that there is significant enrichment in G-quadruplex motifs in gene promoter regions extending up to 1kb upstream of the transcription start sites. These putative promoter G-quadruplex-forming regions are strongly associated with nuclease hypersensitivity sites, and could be involved in gene regulation at the transcriptional level [49].

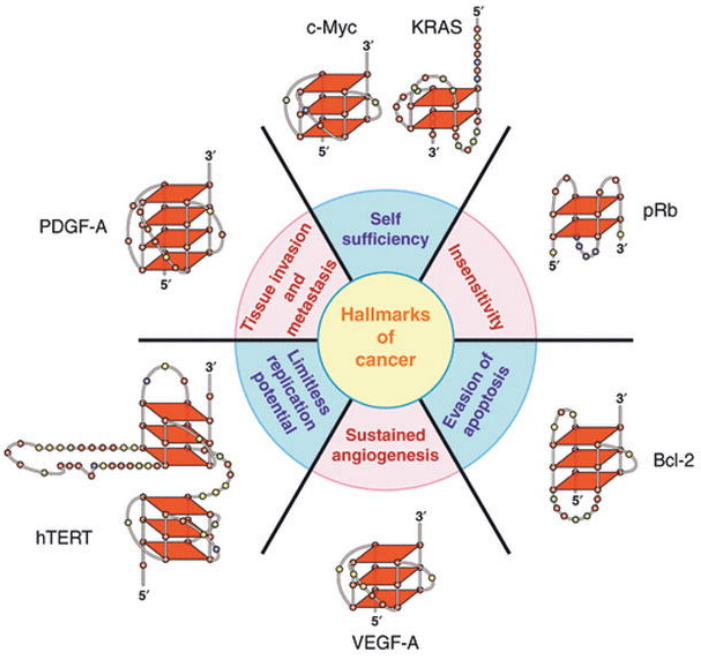


Figure 1.8: **The six hallmarks of cancer shown with the associated G-quadruplexes found in the promoter regions of these genes.** The various G-quadruplexes differ by folding pattern, number of tetrads, loop size and constituent bases. Adapted from [47].

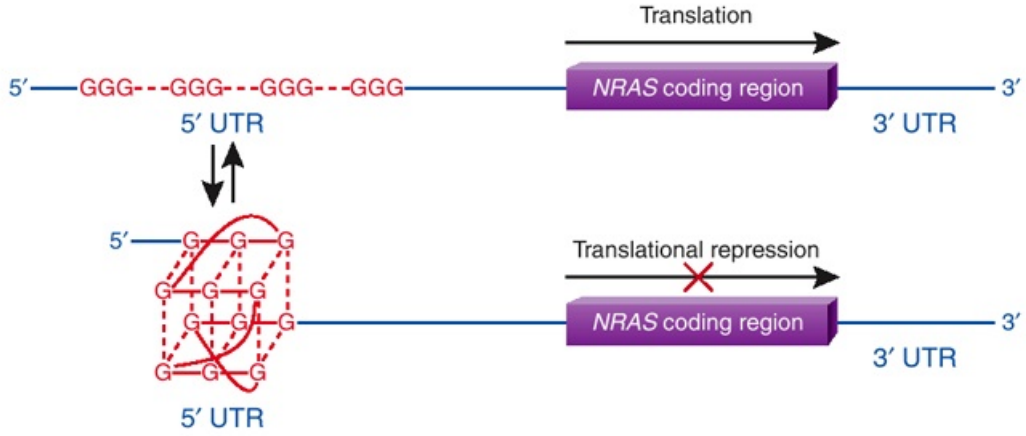


Figure 1.9: **The G-quadruplex structure formed within the 5'UTR of the NRAS mRNA** is proposed to act as a translational repressor element. The G-quadruplex segment is shown in red, and the 5' cap and 3' poly(A) tail are omitted for simplicity. Adapted from [48].

Following transient opening of the duplex (a process necessary for transcription, replication and recombination), these guanine-rich tracts have the potential to form G-quadruplexes. This stimulated investigations on the role of G-quadruplex formation in transcriptional regulation of the promoters (Fig 1.8) of c-myc [50] c-kit [51], bcl-2 [52], VEGF [53] and HIF-1 [54] oncogenes. Another well-known G-rich region is the insulin gene-linked polymorphic region (ILPR), located 363 basepair (pb) upstream of the human insulin gene. Both inter- and intramolecular G-quadruplex formation in the ILPR can influence transcriptional activity of the human insulin gene [55].
The cell-surface-receptor tyrosine kinase. Platelet-derived growth factor receptor β (PDGFR-β), plays an essential role in cellular growth, proliferation, differentiation, and development [56]. The promoter of the human PDGFR-β gene has previously been characterized: the G-quadruplex formed [57] there could be a potential target for drug development in the treatment of cancer, fibrotic disorders [58] and other diseases.

Several studies have demonstrated that the information required for post-transcriptional control is located mainly in the 5' and 3' UTRs of mRNA [59] (Fig 1.9). The secondary structures formed in 5' and 3' UTRs can also serve as regulatory elements by acting as target sites for RNA-binding factors such as proteins or by interacting directly with the translation machinery. Bioinformatic searches have recently identified up to 3000 PQS in the 5'-UTRs of the human gene. A conserved intramolecular G-quadruplex motif within the human NRAS proto-oncogene 5'-UTR, can modulate gene expression at the translational level [60]. These results open opportunities for small molecule identification in order to stabilize 5'-UTR RNA G-quadruplex formation and therefore inhibit translation of oncogenes. We find in the literature other oncogenic 5'-UTR RNAs such as TRF2, fi1, bcl-2 and jun capable of forming G-quadruplexes[14].

Minisatellites & Aptamers

Minisatellites consist of head-to-tail arrays of identical or slightly polymorphic 10-100-bp long motifs. They are present in prokaryote and eukaryote genomes and might constitute as much as 10% of the human genome. It has been shown that they may cause many diseases by influencing gene expression, modifying coding sequences within genes or generating fragile sites [61]. The large number of minisatellite loci, the variety of their nucleotide sequences, and their persistence in the genome suggest that they likely possess biological roles. Some sequences, such as the human CEB25 minisatellite loci are susceptible to form G-quadruplex structures. This 3-quartets parallel G-quadruplex can be formed in long sequence contexts, providing evidence for a pearl-necklace of G-quadruplexes formed by single-stranded CEB25 minisatellite [62].

Aptamers are artificial oligonucleotides (DNA or RNA) selected *in vitro*; they bind a broad range of relevant targets with high affinity and specificity [63]. They consist of relatively small nucleic acid sequence (i.e. usually composed by 15 to 45 bases), that bind to their target with high specificity and with a dissociation constant in the nanomolar to micromolar range. They can be selected against any target such as: small metabolites, peptides, large proteins, nucleic acids and even whole cells.

Moreover, aptamers are easily synthesized and alterable. They are highly used in biotechnology and biological imaging. Some aptamers are characterized for their utility in cancer or viral infections. Indeed aptamers act as inhibitors with great promise, in particular as targeting ligands for the treatment or diagnosis of cancer, HIV and other diseases [64]. They are also characterized by their structure; the folding of these aptamers play an important role in the carried function and their interaction with the targets. Many aptamers fold into G-quadruplex structures. The best known example is the Thrombin Binding Aptamer (TBA) [65, 66]. It forms a monomeric G-quadruplex with two quartets connected by two "TT" loops and a central "TGT" loop. Since then, a lot of G-rich aptamers have been isolated.

Renaud de la Faverie *et al.* analyzed 33 aptamer sequences, extracted from the literature, to determine their topologies. They concluded using different tests, that these aptamers fold into G-quadruplexes [67]. Romanucci V. *et al* designed a bimolecular G-quadruplex aptamers based on Hotoda's sequences d(TG3AG) [68]. These molecules showed significant binding to HIV envelope glycoproteins gp120 and gp41 and represent the first attractive anti-HIV bimolecular G-quadruplexes.

1.3.2 G-quadruplexes and ligands

The examples presented above suggest that the G4s play roles in cells, either at the level of regulation of gene expression, replication, recombination, or division. It is important to develop means to characterize, visualize, induce, stabilize or localize this G4s *in cellulo* or *in vivo*.

There are a number of ligands, proteins and antibodies, characterized by different chemical structures, already reported in the literature for their ability to bind and stabilize the G-quadruplex structures. In the past two decades, hundreds of small molecules with diverse chemical structures and physicochemical properties have been prepared and examined for their abilities to interact with G-quadruplexes [69]. Many of them show *in vitro* and *in vivo* activities, and at least one (Quarfloxin) has entered clinical trials.

Proteins Numerous proteins are able to interact with G4 by: (i) binding to G4 that stabilize the structure, (ii) by unfolding, destabilizing and cleaving the G4, or (iii) inducing G4 formation [70]. The Bloom's and Werner's syndrome proteins are DNA structure-specific helicases. They unwind a variety of very stable G-quadruplexes DNA [71]. The most studied *S.cerevisiae* Pif1 protein can affect the genomic stability of the G-rich minisatellite CEB1, thus providing evidence of their formation *in vivo* [72].

Small molecules Studies of G-quadruplex ligands are now among those at the head of drug discovery, and many databases that include comprehensive information of G-quadruplex ligands are available [73]. The concept that quadruplex-binding small molecules can stabilize telomeric quadruplexes, has been widely used to discover small molecules with anti-cancer activity via indirect inhibition of the telomerase enzyme complex [74, 75].

Most G4 ligands target the invariant part of the structure, the G-quartet, and also possess a large aromatic surface favourable to the interaction with the large hydrophobic surface constituted by the G-quartet. G-quadruplex ligands may be considered as potential anticancer agents because of their telomerase and/or oncogene transcription inhibition potential.

Antibodies Two antibodies (Sty3, Sty49) are able to recognize G4s formed by the telomeric repeats of *Stylonychia lemnae*, with Kd of 125 pM and 3–5 nM respectively [76]. A single-chain variable fragment (scFv1) antibody selected by phage display exhibits a competitive selection to a human parallel intramolecular DNA G-quadruplex with high affinity [77]. The antibody strongly discriminates between G-quadruplex and duplex DNA. More recently, the same team used an antibody to demonstrate the existence of G-quadruplexes in human cells. This will be detailed in the next paragraph.

1.3.3 G-Quadruplexes *in vivo*

Since the first description of G-quadruplexes *in vitro*, key questions have been whether such structures occur *in vivo*, and what their function might be. The first direct evidence for the presence of G4 DNA structures *in vivo* came from studies using G4 DNA-specific antibodies to detect intermolecular structures at ciliate telomeres where their formation and dissolution are cell cycle regulated [76]. Many researcher attempt to localize G-quadruplexes *in vivo*. In 2009 [78], the use of the porphyrine derivative NMM in live *Neisseria gonorrhoeae* showed significantly decreased Antigenic Variation (AV).

Another indirect proof of G-quadruplexes *in vivo* role was performed in the yeast *S. cerevisiae* [42]. This study demonstrated that G4 DNA can contribute to telomere capping which supports the idea that telomere G4 DNA can play a positive role in telomere regulation *in vivo*. The evolutionary conservation of the G4 DNA motif and its association with specific genomic features supports the hypothesis that G4 DNA has *in vivo* functions that are under evolutionary constraint [79]. In addition [80], genome-wide chromatin immunoprecipitation was used to determine the *in vivo* binding sites of the *Saccharomyces cerevisiae* Pif1 DNA helicase, a potent unwinder of G4 structures *in vitro*. The slowing of replication near the Pif1-binding site and the stimulation of DNA breakage suggest that G4 structures form *in vivo* and that they are resolved by Pif1 to prevent replication fork stalling and DNA breakage.

Genetic studies performed on human cells, strengthened the evidence for the existence of G4s *in vivo*. Pyridostatin is a small G4 ligand, which generates DNA damage at specific genomic loci, leading to cell cycle arrest and transcription down-regulation of several genes that contain PQS [81]. Recent studies used antibodies to detect G-quadruplex DNA in mammalian cells. Biffi *et al* [82] have employed a specific antibody to quantitatively visualize DNA G-quadruplex structures in human cells. The BG4 antibody binds with high selectivity and low nanomolar affinity to DNA G-quadruplex structures and does not have any preference to any particular G-quadruplex conformation. The BG4 localization at the chromosomal ends confirms the presence of G4 structures at human telomeres.

Furthermore, the observation of the BG4 dispersed across the chromosomes demonstrates that G4 also forms outside telomeric regions. This study came out with the information that probably most G4 formation occurs during DNA replication and that G4 are modulated dynamically within the cell cycle.

In parallel, another team developed a murine monoclonal antibody called 1H6 [83]. This has also been used to visualize G4 in human and murine cells, providing additional support for the existence of G4 DNA *in vivo*. 1H6 exhibits strong staining in most human and murine cells. This staining indicates the abundance of G4s structure and that the 1H6 antibody is a valuable tool for further studies on the role of G4 DNA.

1.4 Conclusion

G-quadruplexes are prevalent tetra-stranded structures that are formed in both DNA and RNA. They are involved in fundamental biological processes such as transcription and translation. Their structural polymorphism, their presence at different locations within the genome, their association with a number of cancer-related genes and their utilization as potential therapeutic targets, attracted researchers from various areas. Different computational approaches have been developed for specifying the potential of a nucleotide sequences to fold into G-quadruplexes structure. In the next chapter, I will describe the major bioinformatic tools and algorithms developed to predict and analyze sequences with G-quadruplex forming potential.

Chapter 2

Computational detection

Contents

Most biological G4 structures studies have combined *in silico* predictions with biophysical evidence of G4 folding *in vitro*. Computational approaches can nonetheless be a good indicator of the potential of a sequence to form G4 structure, and then, can be used to distinguish sequences that have no potential to form G4-structures from other G4-forming sequences. Different search algorithms have been developed on criteria based on a variety of biophysical experiments. Sequences information alone is not enough to make sure a motif may or may not form a quadruplex, but it is the starting point for the discovery of new G-quadruplexes that should be confirmed using biophysical and biochemical techniques. [84]. The algorithm development started with pattern matching techniques, followed by a sliding window approaches and score calculation. In this chapter, I will bring an overview of the current algorithms used to predict G-quadruplex formation but before this a bioinformatics section should be of interest.

2.1 Bioinformatics

Nowadays, computer science is omnipresent. In biology, it is used in a wide variety of purposes and experimental devices: data archiving, data processing, sequences analysis and predictions etc. Bioinformatics emerged in the 80's as a new discipline where biology, informatics and data processing were fused. Thereby bioinformatics is not only the application of computer science to biology, it is a wholly owned branch of biology.

Current bioinformatics focuses on the study of DNA/RNA sequences and protein folding. It is used for storage or data management and also in the interpretation of such data, whence the data analysis of a sequence might determine the biological function of the gene.

When used to tackle new problems, bioinformatics rely on multiple steps:

- determination of the logical steps required to solve the problem (algorithm)
- identification and structuration of the data (i.e. "objects" used by a program);
- implementation of the algorithm using the appropriate programming languages.

2.1.1 Algorithm

Once the problem is defined and the data involved are identified, one needs to find the calculation mechanism (algorithm) that allows to solve the problem (formula for calculating solutions of an equation, exploring a search tree, etc.).

Algorithm: It consists in a set of logic procedures or formula for solving a problem. The word derives from the name of the mathematician, Mohammed ibn-Musa al-Khwarizmi, who was part of the royal court in Baghdad and who lived from about 780 to 850. His work is the likely source for the word algebra as well[1].

Before validation an algorithm, one should wonder if it is:

- Easy to implement (to write using a programing language)?
- Correct: does it answer the question in all the cases?
- Efficient enough?

Implementation of the algorithm needs an appropriate programming language. Since computers can handle only binary (machine language : succession of only 0 and 1), it is therefore necessary to use a programming language to write legibly the instructions to be executed by the computer; The programming language is the intermediary between the human and the computer. It allows to write in a language close to the machine, but intelligible to the human, whereas, the computer language is a set of consecutive actions that a computer must run.

[1] `http://whatis.techtarget.com/definition/algorithm`

Programming language: Generally, a program is a simple text file, written in any programming language by a specific text editor and called a source file containing lines called source code. This source file is then compiled into machine language to be readable by the processor and executed to solve the problem. The compilation is a phase carried out by the computer itself through another program called compiler.

Different processes and tasks are involved in bioinformatics analysis. Several program have been written for various applications using every available language. There are different levels of programming languages: (i) high-level programing language are easily understood by the user, such as C# , C++, Pascal, Java, Python and Perl. (ii) and the low-level programming language are the one close to the machine language such as assembly.

One can also distinguish *Compiled* and *Interpreted* languages. In the compiled languages, the source code is reduced to a machine-specific instructions before being saved as an executable file. Interpreted languages are saved and executed in the same initial format. The accuracy of each categories and language have been analyzed [85]. The compiled programs generally run faster than the interpreter ones but it is usually easier to develop application in an interpreted environment.

Perl and Python are often called *Script languages* and form a group of intermediate languages. Indeed, when executed, they are compiled in an intermediate representation (without creating an intermediate file) and then interpreted. Both languages have large free libraries, suitable for web scripting and pipeline implementation [86, 87].

2.1.2 Data identification and structuration

Data is a set of qualitative or quantitative variables values, information and knowledge, which are closely related concepts. Data is collected and analyzed to create information suitable for making decisions. Considering a program or an algorithm data should be formatted in a special way. A software is divided into two categories: programs (algorithms) and data can exist in a variety of forms: numbers or text on paper, bytes stored in electronic memory or statement stored in human mind. Data files are the files that store the database information. In bioinformatics, the most known and used data files are FASTA files.

What is a FASTA file? The FASTA format is a text-based format that represents either nucleotide sequences or peptide sequences. The simplicity of FASTA format makes it easy to manipulate using any scripting languages. A sequence record in a FASTA format consists of a single-line description (sequence name) preceded by a greater-than (">") symbol, followed by line(s) of sequence data (bases). The number of sequences in the input data is determined by the number of lines beginning with a ">".

Table 2.1: Different databases, webservers and tools for predicting G-quadruplex motifs. Methods are sorted according to the type of search

Type	Référence	Algorithm	Rules	Devolopement
Pattern Matching Algorithm	[19] [88] [49]	Quadparser /PGS	$\mathbf{G}_x\ \mathbf{N}_{y1}\ \mathbf{G}_x\ \mathbf{N}_{y2}\ \mathbf{G}_x\mathbf{N}_{y3}\ \mathbf{G}_x$ x(3-5) / y(1-7) (later x(2-5)/ y(1-7))	C++
	[89] [90] [91]	QGRS Mapper	$\mathbf{G}_x\ \mathbf{N}_{y1}\ \mathbf{G}_x\ \mathbf{N}_{y2}\ \mathbf{G}_x\mathbf{N}_{y3}\ \mathbf{G}_x$ x $\geq$2 sequences up to 45 bases The maximum length of 30 bases (Max G= 6). Anything accommodating sequences up to 45 bases in length (a single loop of length 0 is allowed)	Web: PHP, Java used for graphics.
		QGRS-H Predictor	Map and analyze conserved QGRS in mRNAs, ncRNAs and other nucleotide sequences e.g. promoter, telomeric and gene flanking regions.	
		QGRS-Conserve	method for calculating motif conservation across exomes	
	[29]	PG4 motif	$\mathbf{G}_x\ \mathbf{N}_{y1}\ \mathbf{G}_x\ \mathbf{N}_{y2}\ \mathbf{G}_x\mathbf{N}_{y3}\ \mathbf{G}_x$ x(2-5) y(1-5) (G4 dans le brin non codant)	
	[92]	Quadfinder	$\mathbf{G}_x\ \mathbf{N}_{y1}\ \mathbf{G}_x\ \mathbf{N}_{y2}\ \mathbf{G}_x\mathbf{N}_{y3}\ \mathbf{G}_x$	Web
	[93]	imGQfinder	Improved Search Algorithm to Find imperfect G-Quadruplexes in Genome Sequences	Web
Sliding window approaches	[94]	G4P Calculator	G-run length>3 number of G-runs per window <4 window length 100 nt;and sliding interval length 20 nt.	C# (Microsoft Windows XP OS).
	[95]	QFP algorithm	GGGNXGGGNYGGGNZGGG where 12+X+Y+Z¡windowsize	Python and available on request
	[96]	ddiQFP		Perl
Score calculation	[97]	scoring system toRNA G4		
Data Base	[98]	GRSDB /GRSDB2	Data Base	Web
	[26]	GRS_UTRdb	relational databases developed with MySQL	
	[99]	QuadBase		MySQL
	[100]	Quadpredict/ QuadDB		

2.2 Pattern-matching algorithms

Mainly three types of computational approaches based on string pattern matching have been used in the literature for analyzing the genomic distribution of PQS. The aim of those algorithms is to search for putative-quadruplexes sequences that follow the motif of the form equation 2.1

$$\mathbf{G}_x\ \mathbf{N}_{y1}\ \mathbf{G}_x\ \mathbf{N}_{y2}\ \mathbf{G}_x\mathbf{N}_{y3}\ \mathbf{G}_x \tag{2.1}$$

In other words, these algorithms are looking for four runs of x guanines separated by three loops of y length. From 2004, a dozen of algorithms appeared [19, 89, 88, 29, 92, 94, 95, 96, 97] (Table 2.1). They all play with the parameters (x, y) of equation (2.1) and the sequences tested. In this section, the major and most cited algorithms for G4 prediction are described.

2.2.1 Quadparser

Huppert *et al.* [19] developed a simple rule describing sequences that may form G-quadruplexes based on the primary DNA sequence. They later investigated the frequency of these sequences and the distribution of the length of the loops joining the tetrads [88].

These methods specifies a regular expression that the PQS should take and consider:

- The strand stoichiometry by limiting the analysis to intramolecular PQS;
- The tetrad number and focusing only on sequences capable of forming three (and later two) or more G-tetrad stacks;
- The presence of mutations or deletions. In this version the discontinuity in the blocks of guanines (presence of bulges) is not tolerated (mutations and deletions in the human telomeric sequence d(GGTTAG)$_n$ results in the decrease in the stability of this sequence);
- The length of the loops is considered between 1 to 7 nucleotides (later up to 12).

The potential quadruplex sequence is then a sequence with four runs of guanines of more than three bases long, separated by loops of 1 to 7 of any nucleotides (equation 2.2). The authors proposed an algorithm named *Quadparser* written in C. Quadparser can rapidly analyze large amounts of genomic data, in FASTA file format, by searching the pattern and report on the number, position of the identified sequences in the output file.

$$\mathbf{G}_{3+}\ \mathbf{N}_{1,7}\ \mathbf{G}_{3+}\ \mathbf{N}_{1,7}\ \mathbf{G}_{3+}\mathbf{N}_{1,7}\ \mathbf{G}_{3+}$$

N means any nucleotides (ACGT) (2.2)

The number of PQS using the folding rule is then calculated considering that:

- when there is more than one possible quadruplex from one sequence it is counted as one (e.g. c-myc d(AGGGT**GGGG**AGGGT**GGGG**) could form G4 using either the first or the forth G of the two bolded blocks)

- when there are more than four runs of d(GGG), it count the maximum number of G4 that could be formed at any given time using consecutive G-runs (i.e. $d(GGGTTA)_8$ would yield a count of 2)
- both G and C-rich patterns are taken into account, meaning that G-quadruplex could be formed in the complementary strand to that for which the sequence was obtained.

The search for prevalent intramolecular G4 in the human genome by Quadparser [88] demonstrates that it contains a large number (376000) of potential sequences. These sequences may be used for a variety of purposes to examine particular genes or other regions of interest.

2.2.2 QGRS Mapper & QGRS-H Predictor & QGRS-Conserve

QGRS Mapper (Quadruplex forming G-Rich Sequences) is another pattern-matching algorithm [92] that aims to predict the presence of G4 in nucleotides entries according to the formula (equation 2.1). QGRS-mapper is a web-based server that generates detailed information on composition and distribution of PQS in any nucleotides sequence in FASTA format. The program provides options to search the entire NCBI, Gene Entrez, RefSeq GeneBank database in order to retrieve the desired gene/nucleotide sequence for analysis, and to provide a user-friendly interface to define the minimum number of tetrads and a maximum length of the loops. The motif again involves four groups of guanines separated by three loops following the restrictions below:

- at least two tetrad are required;
- the default G4 size is 30 nucleotides considering G-groups of a maximum size of 6 while sequences up to 45 nucleotides can be found;
- arbitrary or specified composition of the loops: the user can specify the composition of the loops by a regular expression (e.g.: T{3,5} $\Rightarrow$ loops of three to five consecutive T),
- at most one of loops is allowed to have a length of zero.

In addition to searching for G4 sequences, the tool provides a score that evaluates PQS for its likelihood to form a stable G-quadruplex. The scoring method is dependent on the users parameters (loops length and composition, sequence length...), based on previous studies and uses the following principles for the score calculation: (i) shorter loops are more common than longer ones, (ii) G-quadruplexes are generally of equal loops size and (iii) the greater the number of guanines, the more stable is the sequence.
The highest possible score is 105 corresponding to $(GGGGGGT)_4$. This score allows also the elimination of all possible overlapping sequences by selecting the motif with the highest score. QGRS-Mapper is a flexible and comprehensive tool. The web-based program, written in PHP and java, takes a nucleotides sequence and analyses it for the presence of putative G4. The sequences are mapped to locations such as promoter and poly(A)assuming the original sequences information are provided by the user otherwise this step is omitted. The score is then assigned to PQS for their ability to form G4 and overlapping sequences are eliminated. Finally the results can be displayed in three forms: GeneView, DataView and GraphicsView.

In 2006, the same group developed the **GRSDB** dataBase (G-Rich Sequences DataBase) [98], a database is built to help the request and presents information on G-quadruplex contained in the requested gene. The relational database is built using MySQL. It stocks information about PQS in some region of interest: transcribed regions of alternatively processed human and mouse genes. The analysis of fully annotated GenBank/RefSeq human and mouse genomic using the QGRS-Mapper program allows to constructs the aim data. The database is a single way to study G-quadruplex forming sequences in the RNA processing sites circumstance. Data on composition and location of mapped quadruplexes could be collected; in addition, it display a comparison of G-quadruplex distribution among all the alternative RNA products with the help of dynamically generated graphics. At that time the database contained information obtained from 1310 human and mouse genes, of which 1188 are alternatively processed. A total of 30 584 introns and 33 816 exons were analyzed, containing a total of 3231 RNA products.These products taken together contain a total of 379 223 putative G-quadruplexes, of which 54 252 are near RNA processing sites [within 120 nts of a splice site or a poly(A) signal].

An updated version of **GRSDB** came up in 2008 in addition to a new one named **GRS_UTRdb** [26].
GRSDB2 contains informations on composition and distribution of PQS mapped in a large number of alternatively processed eukaryotic genes in addition to human and mouse genomes. The database allows complex queries with a wide variety of search fields, including Gene Ontology terms. A main characteristic of **GRSDB2** is its ability to correlates the occurrences of G-quadruplexes with gene ontology terms. The data are displayed in a variety of formats with several additional computational capabilities. The statistics indicates that the database includes 29 288 genes with more than 3 millions QGRS mapped to these genes.
GRS_UTRdb contains information on the composition and distribution of PQS in the 5'- and 3'-UTRs of eukaryotic mRNA sequences. It offers a resource for investigating G-quadruplexes in the untranslated regions of mRNA and contains data for more than 16 000 eukaryotic mRNA, including ∼27 000 QGRS, which have been mapped to the 5'-UTRs.

Identifying conserved regulatory motifs helps validate computations and increase accuracy of predictions. In 2012, the same team [90] proposed **QGRS-Homology predictor**, a tool that can map and analyze conserved PQS in mRNA, ncRNAs, promoter, telomeric and gene flanking regions for predicting homologous G-quadruplex forming sequences in the context of 5'- and 3'-UTRs and CDS sections of pair-wise aligned mRNA sequences (Fig 2.1).

QGRS-Homology predictor uses QGRS-mapper for evaluating G-quadruplexes and calculating the G-score and the homology score. The similarities in the location of the G-quadruplexes on the aligned sequence, the number of tetrads, the loops length and the overall lengths are the determinant criteria for the G-quadruplex homology score calculations for two aligned orthologous sequences. The **QGRS-Homology predictor** is a J2EE web application. The user interface is written in HTML and CSS with a significant reliance on JavaScript and jQuery library. BioJava library is used for accessing the NCBI database to quest sequence and meta-data when the user enters accession or GI number as FASTA data inputs. This application identified new G-quadruplexes conserved in the 5'-UTR of the Bcl2 in the human and mouse genome.

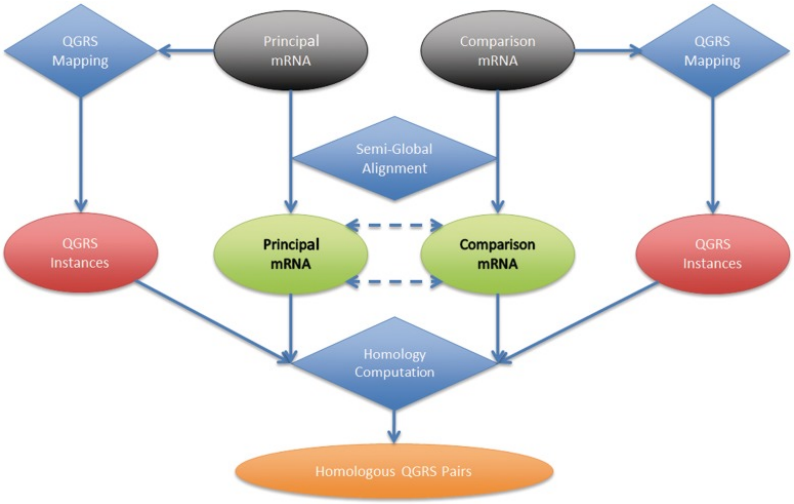


Figure 2.1: **Computational workflow of QGRS-H Predictor.** The QGRS Predictor performs a three-stage computation to produce homology results given two sequences. The results of the QGRS identification stage (performed individually on each sequence) is combined with the results of the semi-global alignment stage to perform the last stage, homology computation. The results of the last stage are filtered according to settings specified in the homology map and presented to the user. Adapted from [90].

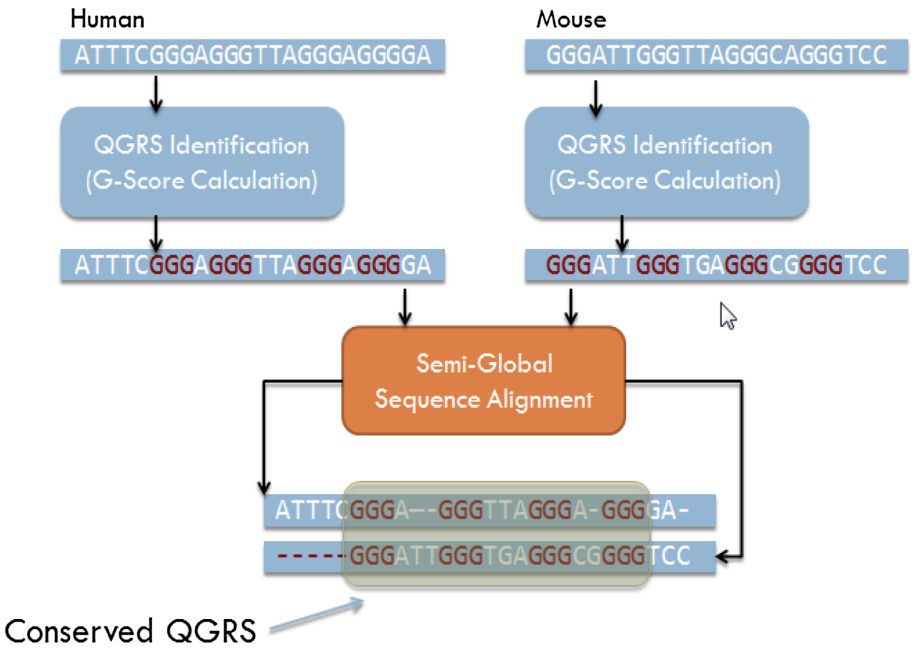


Figure 2.2: **QGRS-Conserve algorithm stages.** Overview of the algorithm stages for QGRS identification (independently) on each mRNA, performing a semi-global alignment on the homologs, and finally evaluating the QGRS pair for conservation. Adapted from [91].

Recently, the same group proposed **QGRS-Conserve**, another computational method for both localizing and calculating G4 motif conservation. It allows the localization and measurement of PQS conservation in the homologous genes. The identification of PQS is based on QGRS-Mapper. Measuring a PQS's conservation across homologous genes is a way of measuring, indirectly, the likelihood that the PQS does indeed form a G-quadruplex. QGRS-Conserve is implemented as a series of three stages (Fig 2.2).

Once the full sequence data for the principal and comparison genes have been obtained, the first stage is to identify all QGRS within each sequence independently. The second stage is a semi-global alignment of the two sequences and the final stage is the calculation of the conservation score. The score of conservation is calculated for each two homologous gene, where 65% of the score was assigned to location similarity, 20% and 10% are for the tetrad number similarity and loop length respectively and finally 5% of the score refer the total PQS length conservation.

With this approaches the authors identified many homologous G4 in the regulatory 5'- and 3'-UTRs of mammalian species and discovered significant differences in their distribution. They suggest tools that can help researchers focus on motifs likely to be of high meaningful interest because of the conservation.

2.2.3 Quadfinder

Quadfinder[2], an online web server for prediction of uni-molecular G-quadruplexes, appeared in 2006 just after QGRS-Mapper. It is designed to be user-friendly, implemented by CGI/Perl and run on Apache HTTP web server v2.0. Based on the equation 2.1, the user have the convenience to set the parameters x and y.

2.3 Sliding window approaches

2.3.1 G4P calculator & QFP algorithm

Eddy and Maizels [94] used, in addition to a sliding window approach to search PQS in various genomes, similar criteria used by J. Huppert *et al.* and V. Scaria *et al.* [19, 92]. They proposed **G4P Calculator** in 2006, an algorithm based on the density of runs of guanines in a sequence. It evaluates runs of guanines in a sliding window that depends on three parameters, (i) k (length of each run of guanines, defaults is 3), (ii) w (window size, defaults is 100 nts) and (iii) s (step or sliding interval, default is 20 nts).

Starting from the beginning of the sequences, the algorithm looks for four runs of guanines of length k separated by at least one nucleotide and shifted by s nucleotides within a window of 100 nts. This approach is more flexible than the regular expression because it limits the length of the candidate but not the length of individual loops. The program is written in C to run on Windows XP operating system.

Similarly to G4P calculator, the **QFP algorithm** based on the pattern equation 2.1, has been implemented in Python [95]. It looks for the presence of four repeats of at least three

[2]No more available online to day.

consecutive guanines each and the length of all the PQS would be less than a window of defined nucleotides length.

It could be expressed as (equation 2.3):

$$\begin{boxed}\mathbf{GGG\ N}_X\ \mathbf{GGG\ N}_Y\ \mathbf{GGG\ N}_Z\ \mathbf{GGG}\\ \mathbf{X, Y, Z} > \mathbf{0}\\ \mathbf{12 + X + Y + Z} < \mathbf{windowsize}\end{boxed} \tag{2.3}$$

The algorithm examines every sliding window that starts with a GGG run, instead of running sliding windows starting at every nucleotide position. This approach gives similar results for a window of 25 nt to the one obtained by Quadparser (3 guanines and loops size from 1 to 7 nts), is more sensitive and allows the authors to survey the distribution of PQS in the *Saccharomyces cerevisiae* genome.

2.3.2 ddiQFP

Previous bioinformatic approaches focused on the identification of genomic sequences having intramolecular G-quadruplexes and several approaches have been developed. These methods yield successfully to the prediction of new G4 validated experimentally, but none of those algorithms appears to identify intermolecular or inter-strand G-quadruplexes.
Wong *et al.* [100] have focused on a novel and complementary methods that search for G4 DNA that assemble from G-runs contained within the two complementary strands of a DNA duplex, which they named ddiQFP (duplex-derived interstrand QFP) and implemented in Perl (see below).

```
# searches for 4 sets of 3-6 consecutive G's or C's with loops up to
#50 nt long, such that the resulting motif is <50 nt long
if ($nongreedy == 1) { $regex =
"(G{3,6}|C{3,6})(.{0,50}?)(G{3,6}|C{3,6})(.{0,50}?)
(G{3,6}|C{3,6})(.{0,50}?)(G{3,6}|C{3,6})"; }
else { $regex =
"(G{3,6}?|C{3,6}?)(.{0,50}?)(G{3,6}?|C{3,6}?)(.{0,50}?)
(G{3,6} ?|C{3,6}?)(.{0,50}?)(G{3,6}?|C{3,6}?) "; }
while ($seq = ? /$regex/g) {
$start = $?[0] + 1;
$motif = substr($seq, $?[0], $ + [0]?$?[0]);
$len = length($motif);
# enforce window size
if ($len <= 50) { print "$start = t$motif" }
# restart search after the 1st run
$nextmotifsearchstart = $+[2];
pos($seq) = $nextmotifsearchstart; }
```

To search for new ddiQFS, a sliding window approach was used based on the search of Hershman *et al.* [95]. The authors defined two main parameters: (i) the length of the G/C runs ($K \geq$ 3nts) and (ii) the window size (w=30, 50, 100 or 200 nts).

The algorithm looks for sequences with: (i) more than six G/C in one G/C run, (ii) when more than six G/C in one run are present, the sequence will be partitioned into two (e.g. GGGGGGGG=G_3GG_4). It eliminates sequences with G_{3+} runs and C_0 runs or vice versa, duplicated motifs and takes the longer motif of the two similar motifs. Finally, it eliminates sequences with loops of length zero. Fourteen different classes of motifs can be distinguished from this approach when examining a single strand of genomic DNA. They could be grouped into nine classes based on the 5' to 3' order of transcription (Table 2.1). With this approach, they found G4 sequences in association with the Pif1 helicase in the yeast and the different classes are distributed in the yeast genome.

2.4 Score calculation

2.4.1 cG/cC score calculator

Beaudoin *et al.* proposed in 2014 an algorithm to search for RNA-G4 [97]. They investigated the formation of G4 located in the 5'- and 3'-UTR and developed a predictive score for G4 folding. The program calculates the RNA sequence's score based on the guanine/cytosine blocs based on the equations bellow (equation 2.4)

The cG score of a sequence s in defined as

$$cG(s) = \sum_{i=1}^{n}(|Gs(i)| * 10 * i)$$

The cC score of a sequence s in defined as

$$cC(s) = \sum_{i=1}^{n}(|Cs(i)| * 10 * i) \tag{2.4}$$

The cG/cC score

$$cG/cCscore = \frac{cGscore}{cCscore}$$

In other words: for a given potential G4 sequences, a value of 10 is attributed to each G or C, and then, a value of 20 and 30 are attributed to each GG/CC and GGG/CCC respectively. The cG and cC score is then the sum of all this attributed values where as the cG/cC score is the fraction $\frac{cGscore}{cCscore}$.

For example: three consecutive guanines generate a cG score equal to 100 because it is counted as three single Gs, where as two consecutive Gs have a total cG score of 40 (equation 2.5).

The cG score of GG sequence

cCscore = 2*(G)*10 + 1*(GG)*20 = 40

The cC score of a GGG sequence

cCscore = 3*(G)*10 + 2*(GG)*20 + 1*(GGG)*30 = 100

(2.5)

The higher the cG/cC score is, the more likely G4 folding and vice versa. According to the authors, this score calculation discriminates between the G4 and the nonG4 sequences with a P-value of 0.0097. Moreover, this method seems to limit the number of false-positive predictions. With more statistical tests, the cG/cC score calculation is both the most sensitive and specific predictor for long RNA transcripts G4 folding.

Finally, the results obtained on a dataset of 14 sequences and the comparison with other algorithm like QGRS-mapper and Mfold, show that the new RNA folding algorithm is a fairly effective predictor of G4 folding.

2.5 Conclusion

These bioinformatics studies have identified several sequences and demonstrated how potential quadruplexes are distributed in the genome. The observed structural diversity of G-quadruplexes offers a different pattern from that sought by these algorithms and the number of quadruplexes predicted may be higher.

Indeed, these studies are based on structural information (pattern matching) and comparative analysis of the location of these sequences [84]. The analysis is performed only on relatively short sequences, which limits the detection of sequences with long loops [16]. The c-myc sequence found within the promoter of that gene has a polymorphism. Indeed, one of the c-myc quadruplexes does not follow the search pattern [20]. Similarly sequences with more than three quartets may also exist [101].

Score-calculation methods are based on the characteristics of the sequence such the length of the loops the number of groups of guanine or the surrounding sequences. Clearly these tools predict new sequences according to feature of existing sequences but their predictions are limited, and a large number of false negatives are generated. The search for a new prediction tools is therefore necessary.

Material and methods

Chapter 3

Experimental detection

Contents

3.1 Products

3.1.1 Buffers

For all spectroscopic experiments, buffers used do not absorb or sparsely absorb light in the studied wavelengths domain (200 to 700 nm) and do not have significant fluorescence properties. The thermal melting experiments require buffers with constant pH through the temperature variations. For this reason, the lithium cacodylic[1] buffer at physiological pH was selected as appropriate buffer for experiments performed at neutral or slightly acidic pH. It was generally used at 10 mM concentration and pH 7.2. For NMR and Thioflavin T experiments, a Phosphate (KPi) and Tris-hydrochloride (tris-HCl) buffer at 10 mM and pH 7.5 were employed.

3.1.2 Oligonucleotides

Oligonucleotides were purchased from Eurogentec (Sereing, Belgium) and delivered in lyophilized form. Stock solutions where prepared in bi-distilled water at 1 mM concentration. Concentrations were determined by absorbance spectroscopy at 260 nm using the Beer-Lambert law and the molar extinction coefficient provided by the manufacturer. The oligonucleotide stocks were stored at $-20°C$.

3.2 Absorbance

Absorbance measurements were performed on Uvikon XL and XS spectrophotometers (Secomam) having a cell changer for 9 samples, 2 references and a temperature probe. This cell holder was thermostatable by an external cryostat. Experiments were performed in quartz cuvets (Hellma, France) with an optical path of 1 cm and a volume of 600 µL.

$$\mathbf{A = \varepsilon.l.c = \log\ (I_0/I)} \tag{3.1}$$

Where:

- ε: the molar extinction coefficient of the absorbing molecule (L. M^{-1}.cm^{-1});
- **l**: length of the optical path of the bushing solution (cm);
- **c**: molecule concentration in the solution (mol.L^{-1});
- $\mathbf{I_0}$: intensity of the incident light;
- **I**: iontensity of the transmitted light.

[1] Li^+ ion has little or no stabilizing effect on the G-quadruplexes

The absorbance A of a species is defined by the Beer-Lambert equation (3.1). Nucleic acids sequences absorb at wavelengths between 200 and around 300 nm with a maximum around 260 nm. The position of maximum is dependent on its sequence structure and base composition: it is close to 255 nm for A and G purine rich sequences, and close to 270 nm for pyrimidine (C, T and U) rich sequences.

3.2.1 Thermal Difference Spectrum (TDS)

Nucleic acid absorption is dependent on both its primary sequence and its structure. A variation in the absorbance may be observed between the structured and unstructured forms of the same oligonucleotide (Fig. 3.1). Thus, by measuring the absorbance as a function of the wavelengths between 220 and 340 nm at two temperatures, it is possible to obtain a differential absorbance curve called TDS (Thermal Difference Spectra). The TDS corresponds to the subtraction of the absorbance spectrum recorded at a high temperature, where the oligonucleotide is completely unfolded, and at low temperature, where the oligonucleotide is completely folded. The technique assumes that these structures are unstructured at high temperature, which is not always the case for G4 as some may still be structured at high temperature. This technique allows to identify a particular G-quadruplex but also other DNA motifs such as B and Z DNA, duplexes, triplexes or the i motif [102]. The typical spectrum of a G-quadruplex consists of two positive peaks to 249 and 270 nm and a negative peak at 295 nm (Fig. 3.1-right).

3.2.2 Isothermal Difference Spectrum (IDS)

Comparable to TDS, Isothermal Difference Specrum (IDS) is also the subtraction of two absorbance spectra. In this case, the experiment is carried out isothermally at 25°C or 4°C [2]. The oligonucleotide previously unstructured [3] is placed in the presence of buffer only (free of cations favorable to the formation of G4) and a first absorbance spectrum is measured. The ion promoting refolding of G-quadruplex is then added (100 mM KCl) and a second spectrum is measured after a 1 hour incubation. The second spectrum is then subtracted from the first after correcting the dilution factor using the equation (3.2).

$$\mathbf{C}_f = \mathbf{C}_i \ (\mathbf{V}_i \ / \ \mathbf{V}_f) \tag{3.2}$$

The shape of the spectra obtained is also based on the nature of the nucleic acid structure studied and can be employed in order to identify the G4 structures. The typical IDS spectrum of a G-quadruplex is related to its TDS.

[2] When the oligonucleotide is not stable

[3] The samples are heated at 90°C for 5 min to unfold all possible structures

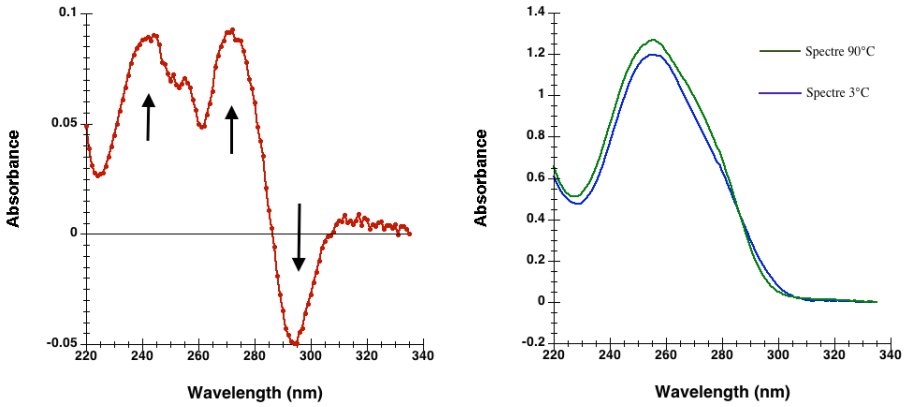


Figure 3.1: **Thermal Difference Spectrum (TDS)**. (right) Subtracting the high-temperature absorbance spectra (green) by those at low temperature (blue) allows to obtain a TDS spectrum (left) with the characteristic peaks of the oligonucleotide studied structure.

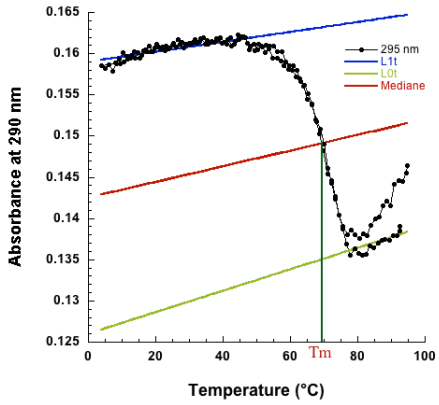


Figure 3.2: **Exemple of Tm determination**.

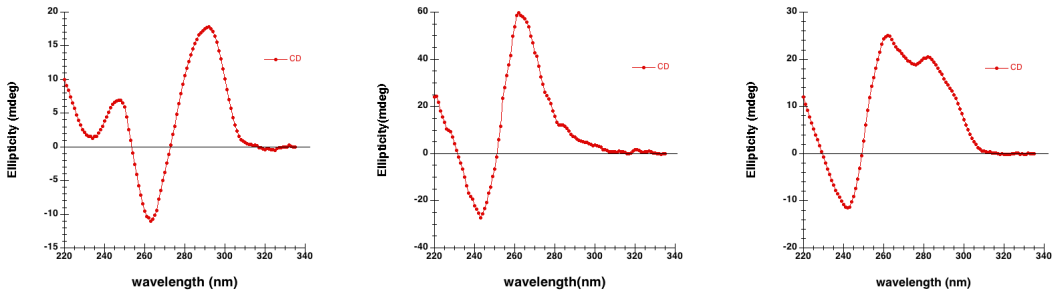


Figure 3.3: **Circular Dichroism wavelength spectra**, (left to rigth) Anti-parallel folding, Parallel folding and signature of a mixed structures.

3.2.3 Thermal melting

The TDS presented above allows to determinate the appropriate wavelengths to monitor the conformational changes induced by a temperature change depending on the structure studied. It is then possible to perform thermal denaturation experiments to monitor the stability of nucleic acids depending on the temperature. It is possible to record a full spectrum at each temperature or record selected wavelengths.

Depending on the structure, the thermal denaturation of a structured nucleic acid may be followed by a hyperchromism (increase in absorbance) or a hypochromism (decrease in absorbance) at a selected wavelength. Regarding the G-quadruplexes, the denaturation of the structure takes the form of a hypochromism at 295 nm. This significant variation determines the temperature of half-dissociation (Tm, Temperature of melting): the temperature at which the molecule is half denatured [103]. The spectrophotometers used have a double beam that allows subtraction, upon acquisition, of the solvent or reference buffer. The manufacture provides three different programs in order to make the appropriate analysis: LabPower, ThermAlys and Julabo.

In this study, bidistilled water was used as a baseline in order to initialize and reset signal between the two carriages. This baseline is generally carried out between 350 and 200 nm at a speed of 200 nm.min^{-1} and with a step of 1 nm using the "LabPower" program. Cryostats are used to thermally regulate sample changers, and establish well-defined temperature gradients (0.2°C.min^{-1}). The ThermAlys program records at regular intervals the absorbance of every sample at 5 different wavelengths (typically 335, 295, 273, 260 and 240 nm). The temperature is recorded throughout the experiment with a probe placed in a reference cuvette. Nucleic acids do not absorb at 335 nm, thus, the absorbance recording at that reference wavelength is subtracted from the measured absorbance at the other wavelengths. The data is then exported and analyzed using KaleidaGraph.

To determine the Tm, it is necessary to draw the tangent of the curve L_1 at low temperature (structured form) and high temperature L_0 (denatured). The Tm corresponds to the intersection between the thermal denaturation curve and the median of the two tangents drawn (Fig. 3.2). The choice of baselines is somewhat arbitrary, as selected by the experimenter, but it does not significantly affect the value of Tm (± 1°C). The structured fraction of nucleic acid at a given temperature (θ_T) should be a number between 0 and 1: $\theta = 0$ for T >> Tm, θ=1 for T << Tm and θ=0.5 for T = Tm.

The conversion of absorbance (A$-T$) to (θ_T) follow the equation (3.3)

$$\theta_T = \mathbf{(L0}_T\mathbf{\text{-}A}_T\mathbf{)/(L0}_T\mathbf{\text{ - }L1}_T\mathbf{)} \tag{3.3}$$

$L0_T$ and $L1_T$ correspond to the baseline values of the unfolded and folded species, respectively.

In this experiment, a cooling is made from 90°C to 4°C followed by a heating from 4°C to 90°C. The respective curves are called renaturation and denaturation profiles; they should be superimposed in the case of structures in thermodynamic equilibrium and absence of artifacts such as evaporation, condensation or sample degradation. If this is not the case, a hysteresis phenomenon is observed due to slow kinetics of denaturation and/or renaturation. This is often the case when bi- or tetra- molecular G-quadruplexes are involved. The association of nucleic acid strands in the G-quadruplexes is often slow; the observed transition does not reflect the thermodynamic equilibrium. It is therefore preferable to call it $T_{1/2}$ rather than Tm.

3.3 Circular Dichroism (CD)

The Circular Dichroism experiments were carried out using a JASCO J810 spectrophotometer, thermostatically controlled by a Peltier. DNA samples concentrations and buffers used were the same as considered for thermal denaturation experiments. Five accumulations were performed and averaged between 335 and 220 nm at a temperature of 20°C. The chosen acquisition parameters are 200 nm / min for the scanning speed with an interval of 1 nm between each reading and 0.5s for the response time. The measurements are performed in quartz cuvettes (Hellma, France) with 1 cm optical path. A baseline corresponding to the buffer is also recorded to be subtracted from the accumulation of spectra after exporting raw data in Kaleidagraph.

CD spectroscopy allows the study of chiral molecules, i.e. molecules having optical activity, and therefore is an attractive method for studying nucleic acids. The circular Dichroism is used to study and evaluate conformation changes in biological macromolecules. Indeed, chiral molecules can change the polarization properties of the radiation passing through them. For a given wavelength, the CD signal is defined as the difference between the absorbance of the left-circular (A_l) polarized light and the right circular (A_r) polarized light (equation 3.4).

$$\begin{boxed} \mathbf{CD = A_l - A_r} \\ \mathbf{A_l = \log(I_0/I_l)} \\ \mathbf{A_r = \log(I_0/I_r)} \end{boxed} \tag{3.4}$$

where I_0 is the incident light fluorescence and I is the outgoing light fluorescence. It is also possible to convert the CD signal (in absorbance unite) to ellipticity θ (millidegree) according to the equation (3.5).

$$\begin{boxed} \boldsymbol{\theta}\ \mathbf{(mdeg) = 32980 * CD} \end{boxed} \tag{3.5}$$

The CD spectrum of nucleic acids results from the transitions due to the couplings between the bases, when they are stacked in a chiral structure. The CD signal reflects the relative orientation among the nucleotides in a DNA structure, and thus it exhibits a particular shape depending on the structure of the oligonucleotide studied. For G4 structures, positive and negative signals at 260 nm to 240 nm, respectively, correspond to a folding in parallel quadruplex; while positive and negative signals at 290 nm to 260 nm, respectively, correspond to an antiparallel fold (Fig. 3.3). CD spectroscopy was employed in order to confirm the results obtained by IDS, TDS and Tm, which were generally recorded in parallel.

3.4 Nuclear Magnetic Resonance (1D NMR)

Nuclear magnetic resonance (NMR) observes the resonance of certain atomic nucleus, such as Hydrogen, when located into a magnetic field. The NMR spectrometer contains a superconducting magnet that generates a stable and homogeneous strong magnetic field (700 MHz 1H, 16.4 Tesla). After insertion into the NMR spectrometer, the magnetic moments of each of the protons in the test sample are aligned with the magnetic field axis. 1D NMR experiments were conducted in two stages: this resonance was disturbed by applying a magnetic field of radio frequency of lower intensity for ten microseconds. Then, this low magnetic field was interrupted, and the return of the magnetization to its equilibrium resonance gave rise to an electromagnetic signal whose frequency is measured.

The frequency of each proton depends on the overall magnetic field but also on the local magnetic field associated with the chemical environment around the proton observed. It is expressed in ppm and each proton has its own chemical shift. The chemical shifts of the various groups of protons that compose one nucleotide are well established. Sugars protons have chemical shifts between 2 and 6 ppm. Regarding the aromatic bases, the protons are found following this distribution: the thymine methyl have a chemical shift between 1 and 2 ppm, the aromatic protons H2, H5, H6 and H8 resonate between 6 and 8 ppm, the protons of the amino group between 9 and 10 ppm and imino protons H1 have chemical shifts between 10 and 14 ppm. In our study we were interested primarily in the imino proton resonances H1 of guanine and thymine. Indeed, these are called proton labile and exchangeable with water and they are visible only when participating in a Hydrogen bond.

The presence of imino peaks between 12 and 14 ppm is characteristic of Watson-Crick base-pair (AT and GC) formation. On the other hand, the presence of peaks between 10 and 12 ppm indicates the presence of imino protons of guanine involved in G-quartets [105]. For example, for a G4 composed of three tetrads, we can count 12 imino peaks, each guanine resulting in an NMR peak. Each G-quadruplex conformation will have a clean NMR signing due to a distribution of imino specific peaks (Fig. 3.4). It is common to see 20 to 30 peaks of different intensities, which reflects the highly polymorphic appearance of G4. Indeed, a G4 sequence can sometimes adopt multiple conformations, and peak intensities reveals the population level of each conformer.

From an experimental point of view, the samples are first denatured at 90 °C in water before

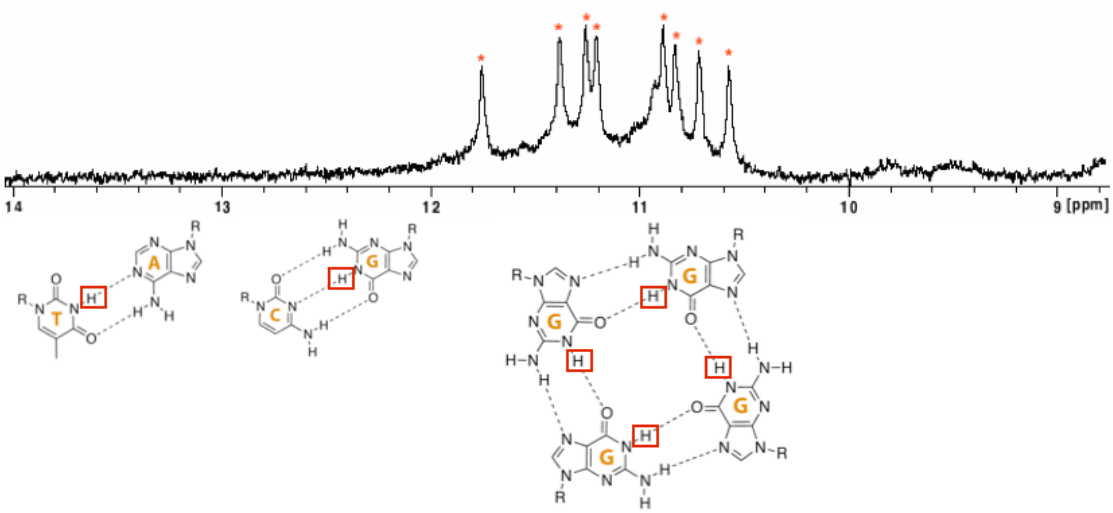


Figure 3.4: **Principle of 1D NMR spectra interpretation**. The imino peaks between 12 and 14 ppm are characteristic of Watson-Crick base-pair AT (13-14 ppm) and GC (12-13 ppm) formation, while the presence of peaks between 10 and 12 ppm indicates the presence of imino protons of guanine involved in G-quartets.

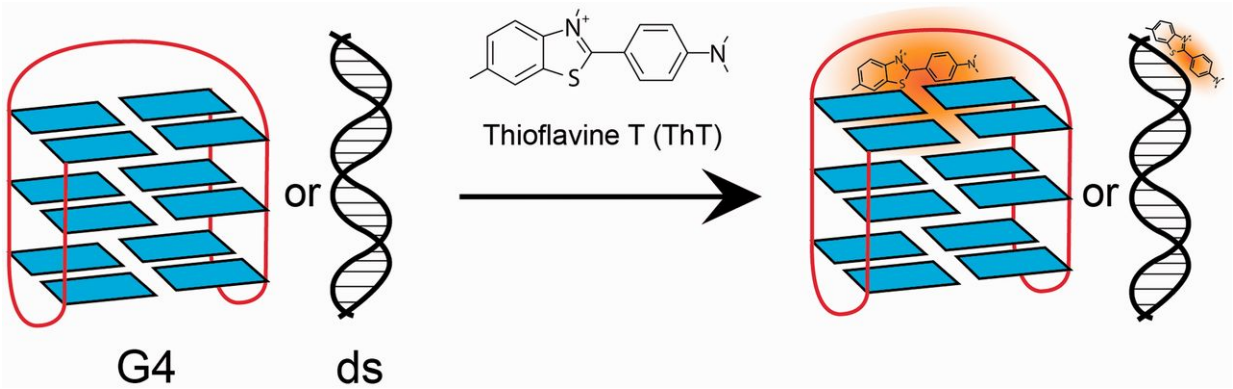


Figure 3.5: **Principle of the ThT assay**. ThT is added to nucleic acid structures preformed in 50 mM Tris-HCl, pH 7.5, and 50 mM KCl. ThT binds specifically to G4, and its fluorescence is enhanced. In contrast, when the oligonucleotide is single- or double-stranded, a lower fluorescence increase is observed. Adapted from [104].

adding 20 mM KPi pH 6.9, 70 mM potassium chloride (KCl) and 10 % of deuterated water (D_2O). The samples, at a strand concentration of around 100 µM, are again heated at 90°C before allowing to cool for 2 hours. The measurements were performed on a Bruker 700 MHz spectrometer at a 4, 25 and/or 38°C. 1D proton NMR experiments were performed with a water suppression such as "jump-return" and a number of scans between 1024 and 2048 depending on the sample concentration for an acquisition time of 10 to 20 minutes.

3.5 Thioflavin T test

The Thioflavin T (ThT) fluorescent probe, initially used in histology for detecting amyloid protein aggregates, can also bind preferentially to a G-quadruplex structure rather than duplexes or single-stranded DNA (Fig. 3.5). ThT has been used for inducing and detecting the human telomeric quadruplex formation due to fluorescence increase compared to DNA duplex [106]. This specificity of ThT for G4 structures was employed to develop a rapid and inexpensive test to determine whether an oligonucleotide folds or not into a G4 structure [104].

The ThT test was performed in a 96-wellplate (Greiner Flat Bottom Black polystyrol) and each oligonucleotide was tested in triplicate. The dye was added to the oligonucleotide solution, and the fluorescence signal was recorded at 490 nm (maximum emission). The fluorescence enhancement is defined as the ratio between ThT fluorescence in the presence or absence of studied oligonucleotides (FI) and the fluorescence of ThT alone (FI_0) after subtraction of the buffer. ThT does not stain all G4 structures equally well. On average the ThT fluorescence enhances more than 60 times in the presence of G4.

In order to analyze whether the oligonucleotides tested fold or not into a G-quadruplex structure, we used c-myc (G4) and ds26 (ssDNA) oligonucleotides as positive and negative controls, respectively, to establish a comparison of the oligonucleotide fluorescence. Preceding the analysis, the oligonucleotides were heated at 90 °C for 5 min at 4 µM concentration in water, next diluted in 100 mM Tris/HCl, pH 7.5, 100 mM KCl and slowly cooled at room temperature for 2 hours. Oligonucleotides and ThT were mixed at 2 and 0.5 µM final concentration, respectively, and each condition was tested in final volume of 20 µL. The fluorescence emission measurement was collected at 490 nm after excitation at 425 nm with a microplates reader (Infinite M1000Pro, Tecan).

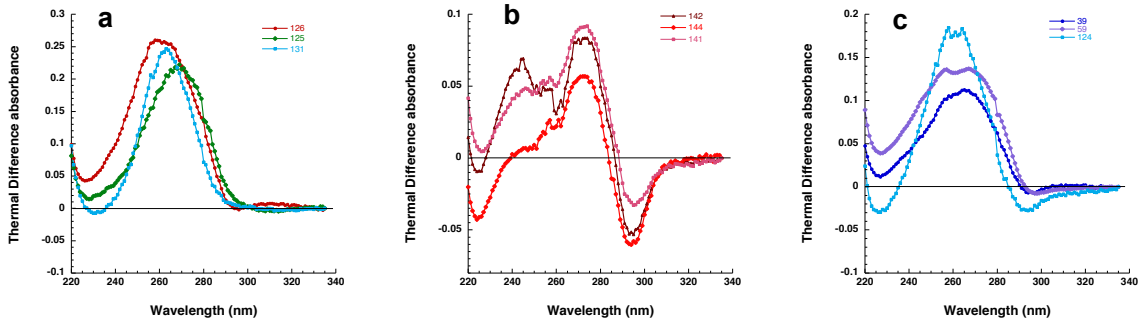


Figure 3.6: **Normalized Thermal Difference Spectra profiles.** The curves are simply the result of the arithmetic difference between high and low temperature absorbance spectra (spectra not show for those examples) of some sequence tested. a) No evidence for G-quadruplex formation b) G-quadruplex sequence with intense negative peak ("Yes"). c) More ambiguous spectra that may suggest G4 formation.

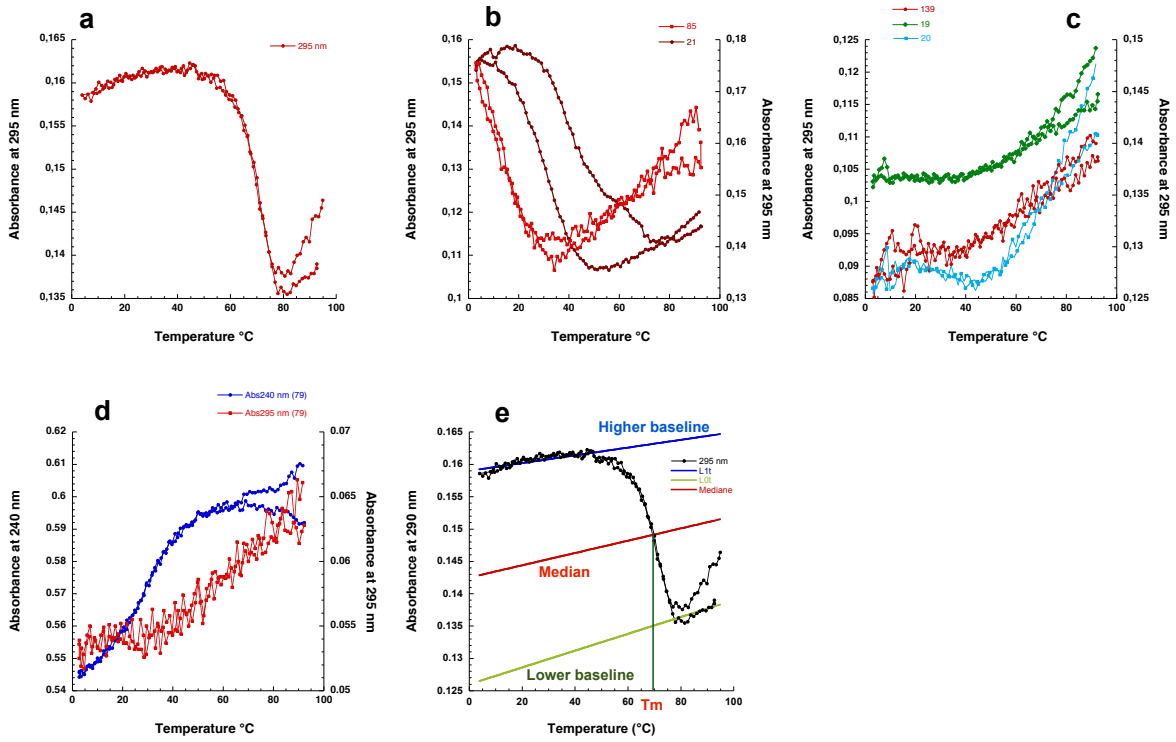


Figure 3.7: **Interpretation of Thermal melting (Tm) profiles.** Measurement of melting profiles enables a characterization of the stability and kinetic of the G-quadruplex association and dissociation. a and b) Profiles of quadruplex forming sequences(a hysteresis is formed for oligo 21 black curves); c and d) Sequences that do not form G-quadruplexes and e) Graphical determination of Tm.

3.6 Biophysical evaluation

Some of the sequences gave ambiguous results. To rationalize our conclusions, we formalized our interpretation of the results as follows:

3.6.1 Interpretation of the Thermal difference spectra

TDS represents the spectral difference between the unfolded and the folded form, provided that the structure can be unfolded at hight temperature. Each type of nucleic acid structure has a specific TDS shape. The G-quadruplexes's fingerprint TDS is characterized by a negative peak at 295 and positive peaks at 240 nm and 270 nm (Fig. 3.6-b) [102].
The more intense the negative peak, the higher the probability to form G-quadruplexes. Depending on the relative intensity of this peak we tagged with "Yes" an intense peak and "Yes (-)" a less intense one (Fig. 3.6-c). The TDS spectrum is not sufficient to prove quadruplex formation and it is always required to interpret this together with data obtained from others methods such IDS, Tm, CD and NMR. The absence of a negative peak at 295 nm might mean that: (i) no structure is formed (ii) a different structure is formed or (iii) the G4 is stable and does not melt (Fig. 3.6-a).

3.6.2 Interpretation of the Thermal melting transition

The thermal stability of the different oligonucleotides was characterized in heating/cooling experiments by recording the UV absorbance at 240 and 295 nm as a function of temperature. We observe a hypochromic shift [103] (decrease of absorbance) at 295 nm upon the melting of G-quadruplexes (Fig.3.7-a/ e). The melting and annealing curves may superimpose or show a hysteresis and the thermal melting temperature may vary from 10 to more than 60°C depending on the sample (Fig. 3.7-a/ b). The sequences tagged with a "Yes" form a G-quadruplex, which can be stable or not according to the Tm. The sequences tagged with "No" either do not form G-quadruplex or G4 is so stable it resists thermal melting (Fig. 3.7-c/ d). Some sequences exhibit an "Inconclusive" profile. Note that the melting of other structures (triplexes, i-motif, etc.) are associated with a hypochromism at 295 nm: an inverted transition at this wavelength is not *per se* sufficient to conclude a quadruplex is formed. Finding that this transition depends on the nature of the cation (K^+ vs Li^+ for example) is a much stronger piece of evidence.

3.6.3 Interpretation of the Circular Dichroism spectra

CD spectroscopy is a biophysical technique widely employed for the validation of quadruplex folded structures [107]. An anti-parallel G-quadruplex form displays two positive maxima around 245 and 295 nm, and a negative minimum around 260 nm (Fig. 3.8-seq4). In contrast, the parallel form exhibits just a single maximum around 260 nm together with the 240 nm negative minimum (Fig. 3.8-seq 62). If the CD profiles of the oligonucleotide follow one of those profiles we indicate the topology "Parallel" or "Antiparallel".

Some sequences display two positive maxima around 260 and 295 nm, and a negative minimum around 240 nm, this mixed spectrum could indicate the co-existence of parallel and antiparallel topologies in solution [108] (Fig. 3.8-seq18, 59), or a hybrid structure. For that type of spectrum we conclude ("Mixture of structure"). For the other CD signature we conclude ("No") (Fig. 3.8-seq 39, 69).

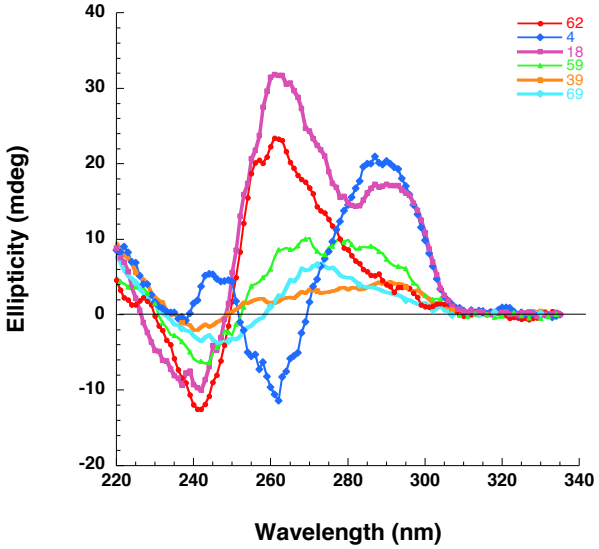


Figure 3.8: **Circular Dichroism spectra of some sequences characterized in this study**. CD spectra were recorded at 20°C. Positive peaks at 260 nm indicate a parallel folded quadruplex, whereas peaks at 295 nm indicate an anti-parallel fold.

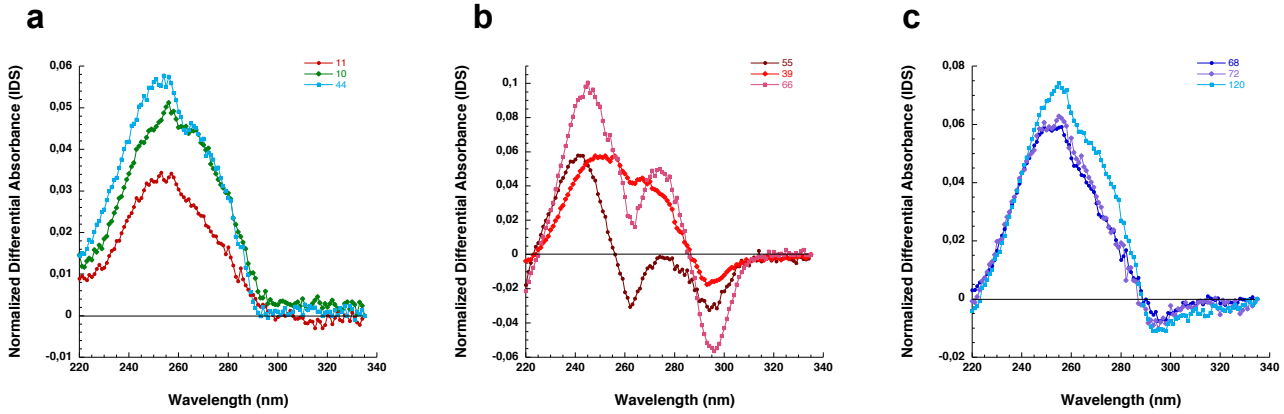


Figure 3.9: **Interpretation of the Isothermal Difference Spectra.** Normalized Isothermal difference spectra results from the difference between the absorbance recorded at 25°C before and after annealing. Data were normalized. a) No G-quadruplex formation b) G-quadruplex sequence with intense negative peak ("Yes"). c) ambiguous spectra with a less intense negative peak ("Yes (-)").

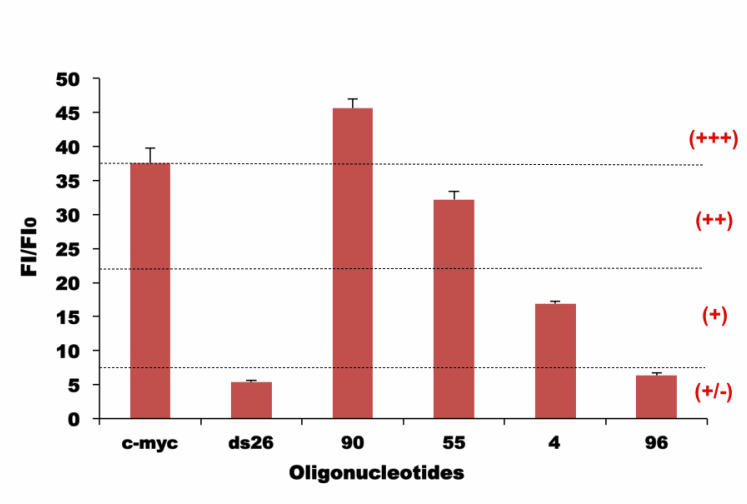


Figure 3.10: **Interpretation of the Thioflavin T test.** Bar graph of fluorescence enhancement of ThT in the presence of a Quadruplex (c-myc), a duplex (ds26) and various oligonucleotides. Error bars correspond to S.D.

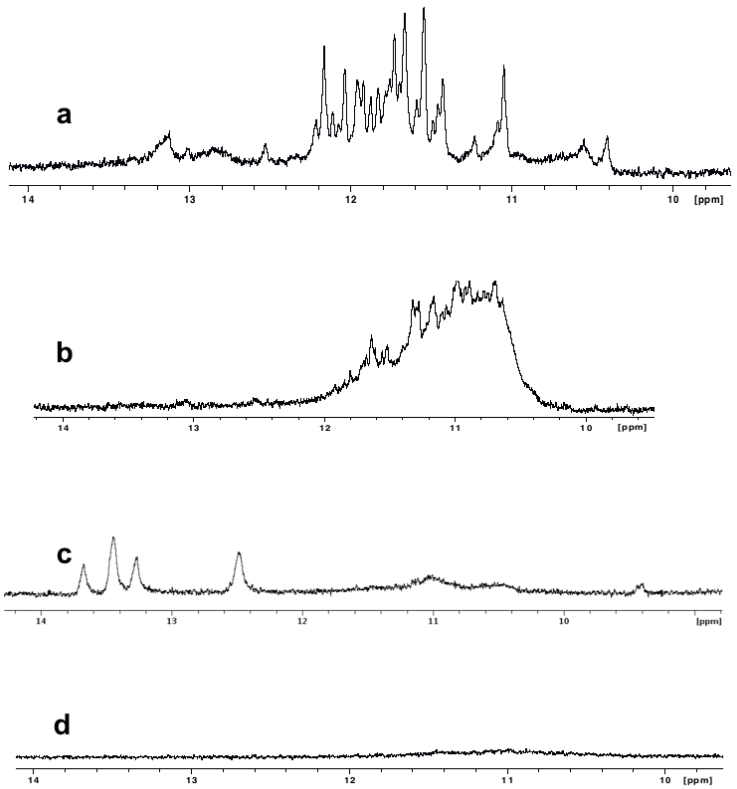


Figure 3.11: **Representative NMR imino proton spectrum of some tested sequence.** The presence of peaks between 10 and 12ppm suggests G4 formation (a,b), the peaks at 12-14 ppm indicate Watson-Crick base pairs (c), no signal is recorded in the absence of G-quadruplex formation (d).

3.6.4 Interpretation of the Isothermal Difference Spectra

The IDS of all the sequences are recorded at 25°C; they were obtained by taking the difference between the absorbance spectra from unfolded (in the absence of salt) and folded (after addition of 100 mM KCl) oligonucleotides [109] and correction of the dilution effects (KCl stock concentration at 3 M). IDS is preferred over TDS as quadruplexes that are highly thermally stable do not unfold at high temperature, and TDS is not informative for these sequences. Similarly to TDS, a negative peak at 295 nm is observed and the intensity of this peak depend on the oligonucleotide.

Depending on the intensity of this peak (very intense (Fig3.9-b) and less intense (Fig3.9-c)) we conclude "Yes", "Yes (-)" respectively. The absence of this peak reflects the absence of the G-quadruplex formation (Fig3.9-a).

3.6.5 Interpretation of the Thioflavin T test

In the presence of G-quadruplexes (we used c-myc as a positive control) Thioflavin T (ThT) emits a strong fluorescence, whereas this enhancement is far more limited in the presence of a control duplex (ds26) or single-strands [104]. We tested our sequences with ThT with a fluorescence plate reader. We compared the fluorescence of ThT in the presence of the sequence tested and the fluorescence in the presence of c-myc and ds26 and chose the following convention:

- if the emission is higher or equal to the c-myc we define the signal as "+++";
- if the emission is in the same range (or lower) as ds26 we mark it as"+/-".
- whenever the signal lies between these two extreme, we label it as "++" or "+" (Fig. 3.10).

3.6.6 Interpretation of the NMR Spectra

For all the sequences ^1H nuclear magnetic resonance (NMR) spectra were recorded at 298K. The presence of peaks between 10 and 12 ppm is consistent with imino protons bound by Hoogsteen Hydrogen bonds and presume G-quadruplex formation (Fig. 3.11-a). The absence of individual discrete peaks in the 10-12 ppm range is consistent with conformational heterogeneity (Fig. 3.11-b-c)

In contrast, the imino protons of Watson-Crick base pairs typically appear at 12-14 ppm, this involves the presence of competition between G4 and duplex formation (Fig. 3.11-c). The absence of imino protons in this region shows the absence of G-quadruplex formation (Fig. 3.11-d).

3.7 Conclusion

Taken individually, none of those methods gave a satisfactory answer in all cases (IDS and 1D-NMR being the most reliable techniques), arguing for the choice of combination of techniques. We therefore based our interpretation on the data generated by CD, IDS, TDS, Tm, ThT and NMR.

Development and validation of a new algorithm: G4-Hunter

Chapter 4

Reevaluation of quadruplex propensity with G4-Hunter

Contents

Several bioinformatic approaches have been developed in order to predict new G-quadruplex-forming sequences. These methods are based on sequence-pattern, sliding window approaches and score calculation. In the meantime, a number of experimental studies have established G4 formation for sequences that escape these predictions (false negatives) while a more limited number of articles reported sequences predicted as G4-forming and still unable to form G-quadruplexes *in vitro* (false positives). Furthermore, most of these algorithms report a binary (yes/no) answer by matching a given pattern, which prevents any quantitative analysis that would have allowed putative correlation of a given quadruplex 'strength' metric with other genomic or functional parameters.
To overcome these limitations, we chose to develop a different algorithm, G4Hunter, that would take into account G-richness and G-skewness of a given sequence and give a score (quadruplex propensity) as an output. To validate this model, we decided to experimentally validate this algorithm using an unprecedented large number of sequences. We applied a combination of biophysical methods to accurately assess *in vitro* quadruplex formation.

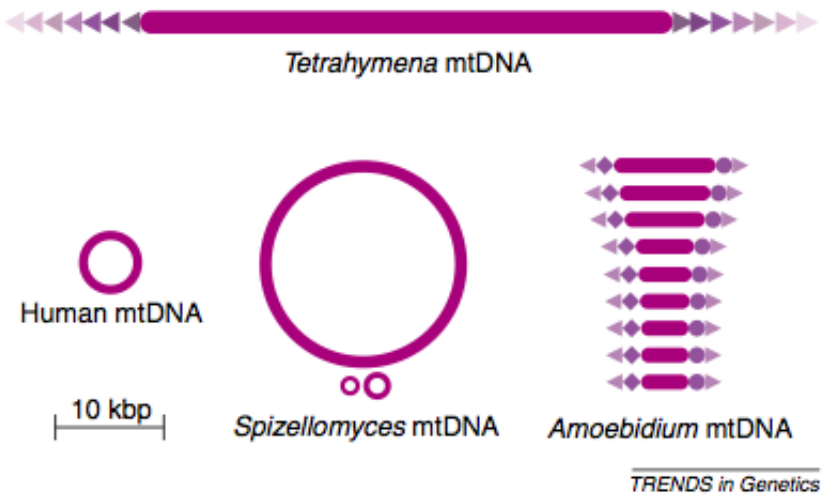


Figure 4.1: **Mitochondrial genome architectures**. Mitochondrial DNA molecules are mostly circular-supercoiled in humans, and linear-monomeric in the ciliate *Tetrahymena pyriformis* [110]. In the fungus *Spizellomyces punctatus*, mtDNA consists of three types of circular-mapping molecules [111], whereas several hundred types of linear-monomeric molecules comprise the mitochondrial genome in the ichthyos-porean protist *Amoebidium parasiticum* [112]. Triangles represent terminal repeat motifs of 32 bp in *Tetrahymena pyriformis* mtDNA and 40 bp in *Amoebidium parasiticum* mtDNA. Filled circles and diamonds represent sub-terminal repeats, which are100-bp and 65-bp long, respectively, in *A. parasiticum* mtDNA. Adapted from [113].

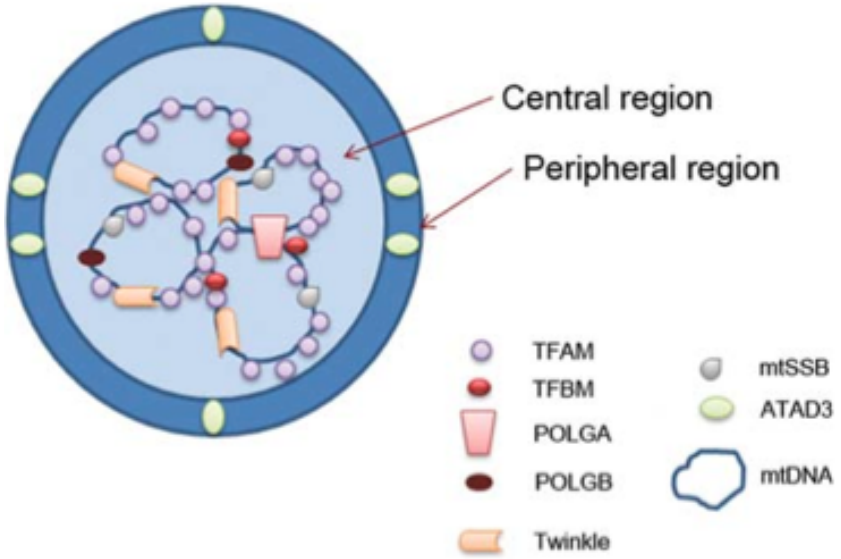


Figure 4.2: **A structural model of the mitochondrial nucleoid**. mtDNA is bound to protein particles forming a sphere, consisting of two zones; central and peripheral. The proteins involved in mtDNA replication (TFAM, mitochondrial transcription factor B-FBM, POLGA, POLGB, MTSSB and Twinkle) are located in the central region, whereas proteins that have an indirect relationship to mtDNA (e.g. ATAD3) are arranged peripherally. Adapted from [114].

We validated this algorithm on a short complete genome, the human mitochondria (16 Kilobase (kb)), because of its relatively high GC content and GC skewness as well as the biological relevance of QFS near instability hotspots [115, 97]. We then applied the algorithm on a series of complete genomes, including the human genome (release EF184640.1). Our results lead to the conclusion that the number of G4-prone sequences in the human genome must be very significantly re-evaluated (by a factor of 2 to 10).

4.1 Mitochondrial genome

Mitochondria are a double-membraned intracellular organelles present within the cytoplasm of nearly all eukaryotic cells. Mitochondria are highly dynamic. Their genetic function is well conserved and fundamental for maintaining cellular homeostasis [116]. Additionally, they have a vital role in calcium signaling and storage, respiration and/or oxidative phosphorylation, metabolite synthesis and apoptosis [117]. The subcellular compartment of the eukaryotic cell use electron transport coupled with oxidative phosphorylation to generate ATP. Although they have a central role in energy transduction, the vertebrate mitochondria are also active regulators of apoptosis and play a central role in metabolism, disease and aging.

The endosymbiotic theory is largely used to explain mitochondrial genetic system [111]. Based on genes located in the mitochondrial genome, phylogenetic analyses indicate that mitochondrial genes were developed from an α-proteobacteria ancestor. The presence of orthologous genes in the mitochondrial genome in some species and in the nuclear genome of other species proves the transfer of bacterial genes from the mitochondrial to the nuclear genome [118].

Mitochondrial genetics differ from Mendelian genetics as uniparental inheritance and cellular polyploidy are some of mitochondrial genetic particularities. Mitochondrial DNA (mtDNA) encodes for key subunits of the electron transport chain and generally used to study molecular phylogenetics. Within human cells, mtDNA is the only constitutive extra chromosomal DNA. Since the first complete sequence of the human mtDNA was published in 1981 by Anderson *et al.* [119], the interest in the mitochondrial genome and its role in the human evolution and disease has dramatically increased.

4.1.1 Structure

Mitochondria exist in various topographical forms, which reflect the type and functions of the cells they serve and thrive off. Mitochondria are the main site of extra chromosomal DNA within the cell (except chloroplasts in plants), and are under the genetic control of both the nuclear DNA and mitochondrial genome [120].

The mtDNA is a double-stranded closed-circular molecule. Linear mapping have also been detected in some apicomplexa (*Plasmodium falciparum*), fungi and several cnidarian animals [121]. The mitochondrial genome typically consists of a single chromosome (Fig. 4.1) except some cases such as *Spizellomyces punctatus* [122], *Globodera* and the protistan *Amoebidium*

parasiticum (which involves a complex set of hundreds of linear molecules) [112]. Similarly, the kinetoplastid protists (e.g. *Trypanosoma*) mitochondria called the kinetoplast contains up to few dozen gene-coding maxicircles and up to several thousand minicircles that are involved in editing mitochondria mRNAs [123].

The size of mtDNA in most Eukaryotes ranges from 15 to 60 kbs; however, there are some exceptions and much larger mitochondrial genomes have been found. These are due to duplication of some portions of the mtDNA rather than variation in gene content [124]. Among the organisms whose mtDNA has been sequenced, we find apicomplexan protists *Plasmodium sp* with a minuscule mitochondrial genome of 6 kbp, and rice (*Oryza saliva*) whose mtDNA (490 kbp) is 80 times larger than the *Plasmodium* [125]. The number of mitochondrial nucleoids per mitochondrion ranges from 1 to 10 (Fig.4.2). Recently quantitative analysis of the size and mtDNA content of nucleoids in cultured mammalian cells suggests that an average mitochondrion may contain 5 to 7 mtDNA genomes [126].

The *Homo sapiens* mitochondrial genome is a closed-circular, double-stranded DNA molecule of about 16.6 kbs containing 37 genes. The two strands are distinguished on the basis of G+T base composition, which results in different buoyant densities of each strand ('heavy' and 'light')[1]. Most information is encoded on the heavy (H) strand, which contains genes for two rRNAs, 14 tRNAs, and 12 polypeptides. The light (L) strand codes for eight tRNAs and a single polypeptide. All 13 proteins are constituents of the enzyme complexes of the oxidative phosphorylation system (OXPHOS). The signature form of the mammalian mtDNA is the displacement-loop (D-loop), which contains sequences that are vital for the initiation of both mtDNA replication and transcription [127]. mtDNA displays an exceptional economy of organization. The genes lack introns and except for one regulatory region, intergenetic sequences are absent or limited to a few bases. Both rRNA and tRNA molecules are unusually small. Some of the protein genes are overlapping and, in many cases, part of the termination codons are not encoded but are generated post-transcriptionally by polyadenylation of the mRNAs.

[1] In denaturing caesium chloride gradients

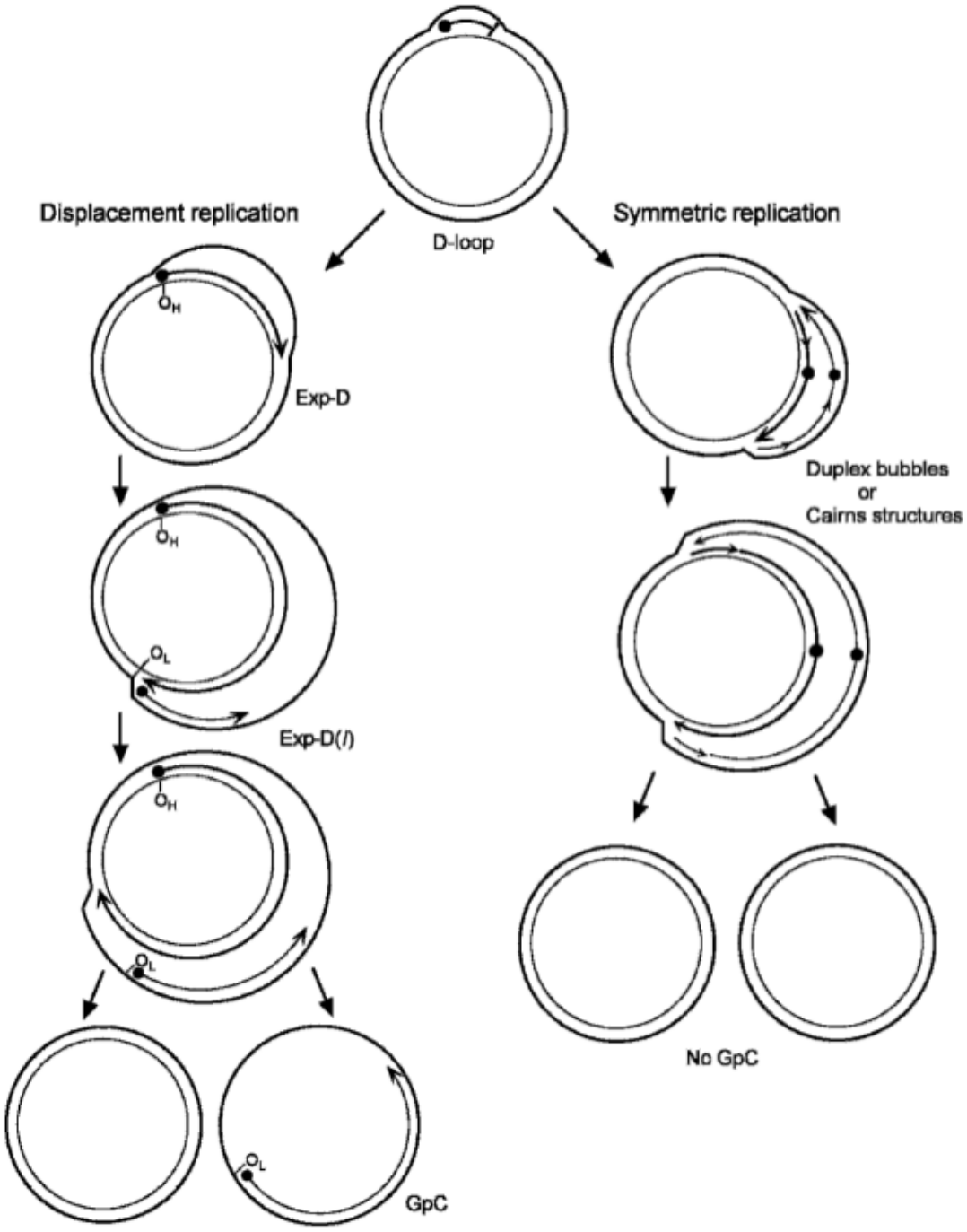


Figure 4.3: **The asymmetric and strand-coupled models of mtDNA replication** . Both models agree on the nature of the simple D-loop form of mtDNA. The displacement-model of replication is shown on the left and proceeds with single-stranded replication of the H-strand with further expansion and displacement of the D-loop. The intermediates are called expanded D-loops (Exp-D). This proceeds until the L-strand origin (OL) is exposed, with subsequent synthesis of the new L-strand in the opposite direction. Those intermediates are termed Exp-D(I). The asymmetry of strand synthesis leaves one segregated daughter molecule with an incompletely synthesized L-strand, called a gapped circle (GpC). The strand-coupled or synchronous model of replication is shown on the right. In this model, there is thought to be a zone of replication initiation within a broad area beyond the simple D-loop. Within this zone, both strands are synthesized bidirectionally as the double-stranded replication forks proceed through the length of the mtDNA. Adapted from [128].

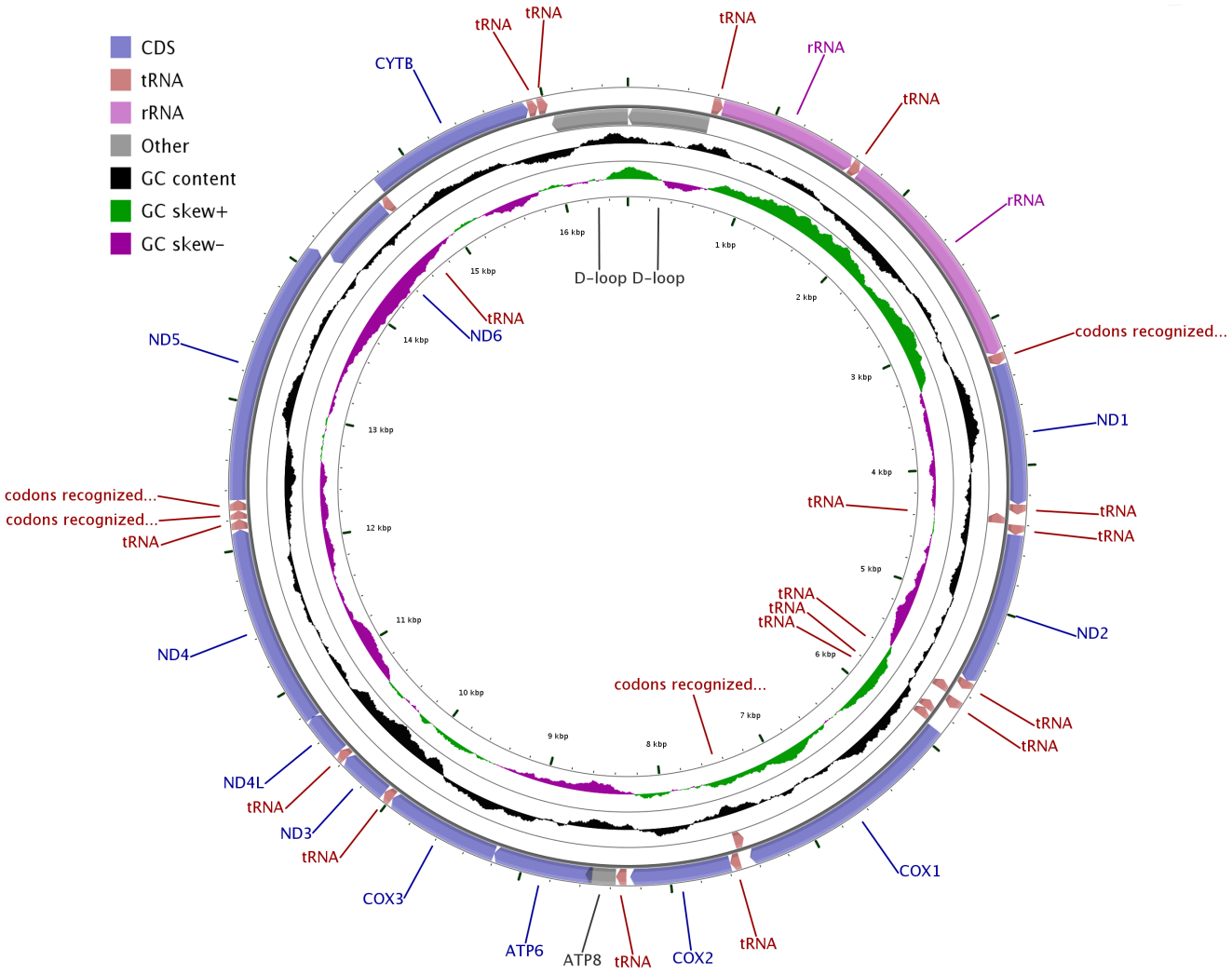


Figure 4.4: **Human mitochondrial genome map.** The human mtDNA genome encodes 13 polypeptides. These include seven subunits of Complex I (ND 1–6 and ND4L), one subunit of Complex III (CytB), three subunits of Complex IV (COX I to III) and two subunits of Complex V (ATPase6 and ATPase8). It also encodes two rRNAs (12S and 16S) and 22 tRNAs. The main control region is the D-loop which contains the H-strand promoter region (HSP), the LSP region and the origin of H-strand replication (OH). The second control region consists of only 30 bp and is located between ND2 and COXI and is the site of the origin of L-strand replication (OL). tRNA and rRNA-coding genes are shown inside and outside the circles. The forward and reverse DNA strands are shown in clockwise and anticlockwise orientation, respectively the second black peaked circle represents the GC content, followed by green and pink peaked circle representing GC skiw + and GC skew - respectively. The inner circle shows the size markers in kbp in clockwise orientation. This figures was generated using the CGview server. `http://bioinformatics.org/cgview/`

4.1.2 Replication and transcription

Mitochondrial replication is mediated by several nuclear-encoded transcription and replication factors, which along with mtDNA are packed in the mitochondrial nucleoid of ∼70 nm in diameter.

Replication of mammalian mtDNA is described by the asymmetric and the coupled leading- and lagging-strand (symmetric) models [128]. The asymmetric model (Fig. 4.3) has been the traditional approach to understanding how mtDNA replication is mediated [129]. It is initiated from the origin of H-strand replication (O_H), located within the D-Loop region, where mitochondrial transcription factor A (TMFA) binds to the enhancer of the light-strand promoter (LSP) and induces structural changes that expose the promoter region to the mitochondrial-specific RNA polymerase. This allows an RNA primer to be generated, which is used to initiate mtDNA replication. mtDNA replication then continues two-thirds round the genome to the origin of L-strand replication (O_L). L-strand synthesis then proceeds in the opposite direction. The coupled leading- and lagging-strand replication (symmetric) model proposes that both H- and L-strands are replicated bidirectionally from the same initiation cluster site [127] (Fig. 4.3). This mechanism might occur in addition to the asymmetric model, thus for repopulating mtDNA. However, to date, there has been little resolution as to which model is the most appropriate with entrenched views being expressed by the two opposing groups.

In some mtDNA, all genes are transcribed from one strand, whereas in the others, the genes are distributed between the two strands. The human mitochondrial genome contains two major promoters referred to as the light strand promoter (LSP) and the two heavy strand promoters (HSP1 & HSP2). Transcription initiated at LSP and HSP2 located in the D-loop. It generates polycistronic transcripts which are processed to produce individual mRNA and tRNA molecules [130, 131]. The HSP1 produces only two rRNAs (12S and 16S) and two mt-tRNAS ($tRNA^{Phe}$ and $tRNAs^{Val}$). Transcription of mtDNA depends on few nucleus-encoded proteins: a single RNA polymerase (POLRMT), an activation (Tfam) and a termination factor (mTERF) as well as auxiliary factors (such as TFB1M, TFB2M) necessary for promoter recognition. This simple system can consider the bidirectional transcription of mtDNA from divergent promoters and key termination events controlling the rRNA/mRNA ratio [132].

4.1.3 Function

The coding capacity of mtDNA ranges from nearly 100 genes in flagellates [133] to only five in *Plasmodium*, with the average across eukaryotes being 40-50 genes.
As explained before, the typical human mtDNA genome contains 37 genes encoding for 13 protein subunits of oxidative phosphorylation enzymes, two rRNAs of the mitochondrial ribosome and 22 tRNAs necessary for the translation of proteins encodes by mtDNA. The remaining protein subunits, which complete the OXPHOS, are encoded in the nuclear DNA, synthesized in cytosol and imported into mitochondria (Fig. 4.4). The variation of the gene content has been shown in nematodes which lack A8 (ATP synthase subunits

8), in bivalves which both lacks A8 and contains an extra tRNA, and cnidarians, which have lost nearly all tRNA genes and gained one or two additional genes. All the 37 genes found in the animal mtDNA have homologs in the mtDNA of plants, fungus and/or protist [134].

Mitochondria are highly dynamic. Their genetic function is well conserved and fundamental for maintaining the cellular homeostasis [116]. Additionally, they have a vital role in calcium signalling and storage, respiration and/or oxidative phosphorylation, metabolite synthesis and apoptosis [117] The sub-cellular compartment of the eukaryotic cell use electron transport coupled with oxidative phosphorylation to generate ATP. Although they have a central role in energy transduction the mitochondria in vertebrates are also an active regulators of apoptosis and play a central role in metabolism, disease and aging.

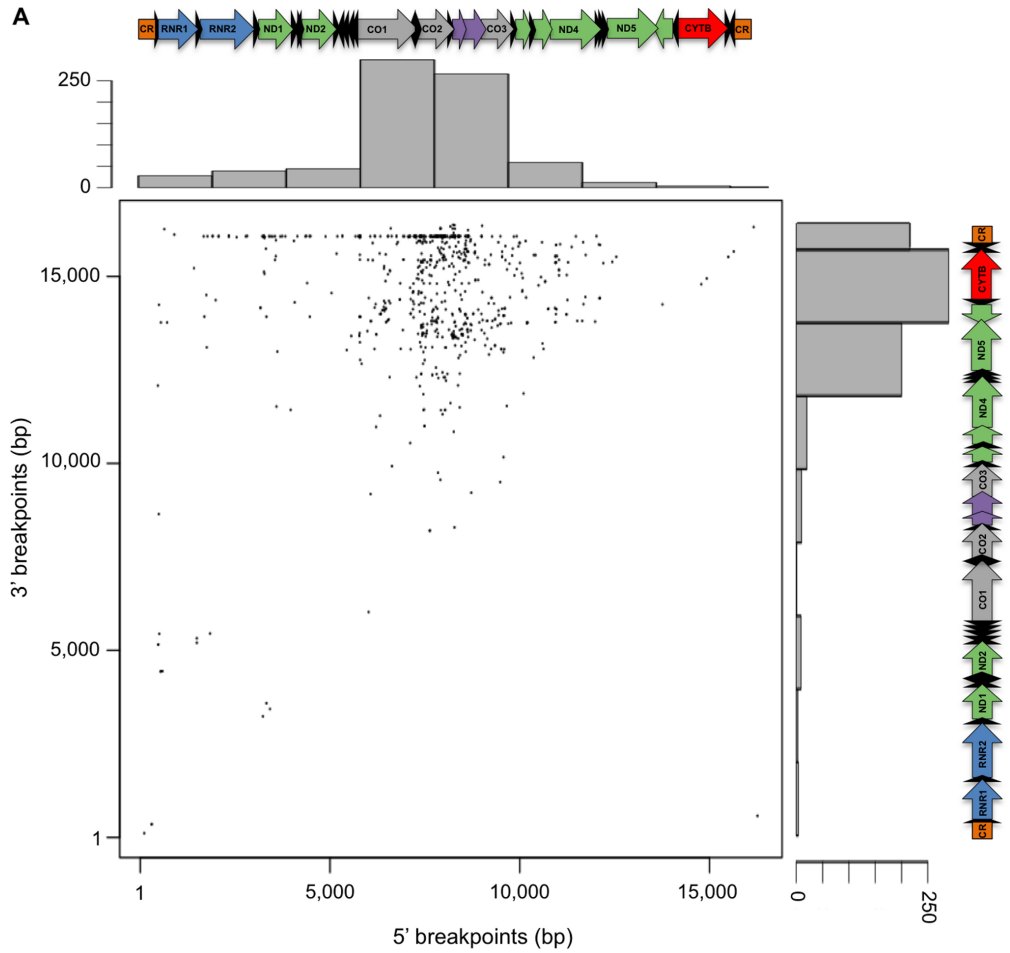


Figure 4.5: **Human mitochondrial deletion spectra.** (A) Distribution of the 5'- and 3' positions corresponding to 1,508 breakpoints, as well as corresponding histograms and positions along the mitochondrial genome. CR-control region; RNR-Ribosomal RNA; ND-NADH dehydrogenase; CO-cytochrome oxidase; CYTB-Cytochrome B. Black arrows correspond to the 22 tRNA genes. Adapted from [135].

4.2 G-quadruplexes & Human mitochondrial DNA

4.2.1 Human mitochondrial genome

Mitochondrial DNA (mtDNA) occurs in nucleoids in multiple copies. Although a mtDNA repair system does exist, the mitochondrial genome has a high mutation rate, 10 to 17-fold higher than that observed in the nuclear DNA [136]. A number of human diseases originate from mtDNA mutations causing mitochondrial disorders and dysfunction. The first pathogenic mtDNA mutations were identified in 1988 and since then over 250 pathogenic mutations have been identified [137]. The most severe mutations are deletions cause respiratory dysfunction. Epidemiological studies demonstrated that single deletions causes about 13-30% of primary mitochondrial diseases. Recent studies have reported that multiple mtDNA deletions may occur in somatic tissue and accumulate with age [115].

mtDNA deletions are prominent in genetic disorders such as Kearns-Sayre Syndrome (KSS), Pearson's Syndrome (PS), and Progressive External Ophthalmoplegia (PEO) as well as various cancers and age-related diseases. The mtDNA deletions breakpoints show a non-random distribution that seems to reflect the mechanism underlining the formation of deletions. Sequences predicted to form stable stem-loop or cruciform structures overlap with only a subset of the total 5'- and 3' deletion breakpoints [138] indicating the potential implication of other non-B form structures in mtDNA instability [139]. In the nucleus, G4 structures can inhibit DNA replication and cause genome instability [139, 72], and PQS have been associated to deletions and duplications in the genomic DNA of cancer cells [140].

There has been an increase interest in nuclear G-quadruplex in recent years; however, researchers have begun to focus on the mitochondrial genome and its ability to form G-quadruplexes only recently.

4.2.2 G-quadruplexes and mitochondria

Zakian *et al.* [79] were the first to predict mitochondrial G-quadruplexes forming sequences in *Saccharomyces cerevisiae* and found that its AT-rich mtDNA had a higher density of G4 motif comparing to nuclear chromosome (0.373 and 0.067 respectively). G4 formation was favorable towards regions that do not encode ORFs, tRNA or rRNA genes. The evolutionary conservation of the mtG4 motifs was not analyzed due to the scarcity of mitochondrial genome of other fungi.

The mitochondrial genome encodes essential components of the respiratory chain and several human diseases are caused by mtDNA deletions. Multiple studies attempted to identify the origin of these deletions. The first features identified were short direct repeats that contain or flank the deletion junction [141] but leave a significant part of these mutations unexplained. Different studies support the hypothesis that the primary cause of mtDNA susceptibility to breakage resides in the formation of non-B DNA conformations [72, 139, 138].

Damas *et al.*[138] collected information about all the available mtDNA deletions and found that the 5'- and 3'-breakpoints are often found around and within COX2 and in the position 16071 respectively (Fig. 4.5-A). They noticed that AT-rich regions around break points might contribute to the formation of single-stranded DNA upon supercoiling and some of those regions are highly structured: hairpins, cloverleaf and cruciform structures. They found out that the genomic regions with breakpoints have a significant higher folding capacity than regions with a lower number of breakpoints.

Previous experiments have shown that non-B DNA acts as a preferential interruption site for DNA polymerases, and might also serve as a recognition motif to the binding of topoisomerases and nucleases, thus permitting DNA breakage and deletion formation. Thereby, formation of non-B DNA conformations (hairpins, cruciform, and cloverleaf) should be considered as important structures for mitochondrial genome rearrangements.

In the nucleus, G4 structures play a role in DNA replication [139] and cause genome instability [72]. Additionally sequences with G4 forming potential (QFP) have been associated with deletions and duplications in the genomic DNA of cancer cells [140]. Another study reported the formation of DNA triple-helices between repeats flanking mtDNA deletions and/or duplications [142]. Little is known however about QFP sequences in the mitochondrial mammalian genome and, considering all of the above results, evaluating the importance and the impact of non-canonical structures on mtDNA instability is a necessity.

The human mitochondrial transcription machinery generates the primers required for LSP replication's initiation. A large fraction of the LSP transcriptions (>65%) were prematurely terminated at the conserved sequence block II (CSBII). The work of Wanrooij *et al.*[143] demonstrated that the CBSII sequence has a strong potential to form G4 structures at the RNA level, which affects the formation of stable RNA primer. An extensive screening of human mtDNA deletions for the presence of G4 structures led to propose that the G4 motifs are likely to cause mtDNA rearrangements, which are at the origin of human genetic disorders such as cancer and aging [135].

Recently, two major studies attempted to localize and associate the mitochondrial DNA deletion breakpoints with G-quadruplex formation [115, 144]. The study of D. W. Dong *et al.* [115] was the first to set up the prevalence of G4 sequences in mammalian mtDNA and reveal their association with mtDNA breakpoints. During replication, the heavy strand is susceptible to form G4 structures. The higher density of G4 structures in the deletion hot spot near the D-loop leads the formation of a single-stranded DNA or double-strand breaks which are known to cause mtDNA deletion formation through a non-homologous end joining (NHEJ) repair mechanism. The QFP sequences may also prompt deletion formation during DNA repair and the association of these structures to the direct repeat pair ($p<10^{-14}$) favors the significant association of G4 with deletion breakpoints. Finally to maintain the mtDNA stability and thus prevent disease, PQF sequences and multiple helicases (PEO1, RECQL4 and PIF1) are susceptible to provide mitochondrial genome stability.

It is also believed that G-quadruplex formation is a probable source of mitochondrial instability by perturbing the progression of the replication machinery.

The computational analysis of S.K.Bharti *et al.* [144] revealed that 5' and 3' deletion breakpoints density was about 1.4 - 7.6 fold greater in G4 compared to tRNA or D-loops. In this study, biophysical approaches were used to demonstrate the *in vitro* folding of some sequences, and suggested that mitochondrial G4 (mtG4) may be a source of mitochondrial genome instability in certain genetic mitochondrial disorders, cancer and aging.

Additionally, a striking feature of the human mitochondrial genome is its GC *skewness* (the non-coding "Heavy" H-strand is G-rich while the complementary "Light" coding L-strand is C-rich), making it a good candidate to strengthen the ability of G4-hunter to predict G4 sequences. We decided to identify G4 forming sequence (G4FS) in the whole mtDNA genome dataset in which we could perform an exhaustive search for G4 propensity.

4.3 Article: Reevaluation of quadruplex propensity with G4-Hunter

Amina Bedrat[1,2], Laurent Lacroix[3,*]& Jean-Louis Mergny[1,2,*]

1. Université de Bordeaux, ARNA Laboratory, F-33000 Bordeaux, France;
2. Inserm U869, IECB, F-33600 Pessac, France;
3. CNRS-Université de Toulouse UMR5099, F-31000, Toulouse, France.
** Corresponding Authors*

Reevaluation of quadruplex propensity with G4Hunter

Amina Bedrat[1,2], Laurent Lacroix[3]* & Jean-Louis Mergny[1,2]*

1. Université de Bordeaux, ARNA Laboratory, F-33000 Bordeaux, France;
2. Inserm U869, IECB, F-33600 Pessac, France;
3. CNRS-Université de Toulouse UMR5099, F-31000, Toulouse, France.

* Authors to whom correspondance may be addressed:
(LL) laurent.lacroix@inserm.fr
(JLM) jean-louis.mergny@inserm.fr

Abstract:

Critical evidence for the biological relevance of G-quadruplexes (G4) has recently arised thanks to seminal studies performed in a variety of organisms. Four-stranded G-quadruplex DNA structures are promising drug targets as these non-canonical structures appear to in be involved in a number of key biological processes. Given the growing interest for G4, accurate tools to predict G-quadruplex propensity of a given DNA or RNA sequence are needed. Several algorithms such as Quadparser allow to predict quadruplex forming propensity. However, a number of studies have established G4 formation for sequences undected by these tools (false negatives) while a more limited number of articles reported sequences obeying the consensus and still unable to form G-quadruplexes *in vitro* (false positives).

In this manuscript, we chose to develop a radically different algorithm, G4Hunter, that takes into account G-*richness* and G-*skewness* of a given sequence and gives a score (quadruplex propensity) as an output. To validate this model, we decided to benchmark this search using an unprecedented large number (500+) of sequences. We first developped a combination of biophysical methods to accurately assess quadruplex formation *in vitro*. We experimentally validated this algorithm on a short complete genome, the human mitochondria (16.6 kb) because of its relatively high GC content and GC skewness as well as the biological relevance of these quadruplexes near instability hotpsots. We then applied the algorithm on a number of species including human, allowing us to conclude that the number of *bona fide* G4-prone sequences in the human genome should be reevaluated very significantly, by a factor of 2 to 10.

Keywords:

Quadruplex; sequence prediction; G-quartet; G4 propensity

Introduction:

Nucleic acids appear more and more than just a passive instruction manual for the cell. Above the information content of nucleic acid bases succession, genomic material spatial organization, bases modifications and chromatin accessibility are important components for cell functioning. Misregulation of these elements and their usage are linked to many cellular dysfunctions. At a smaller scale, the succession of bases responsible for the wording of this information also codes for the ability of a given letter sequence to interact with other letters. This supports the classical double-helix formation for the DNA and its well-known advantages (1) and also the much more complex folding of RNA molecules and their diversity of function. More and more light has been recently shed on alternative or unusual nucleic acids structure as sequences prone to such 'unorthodox' structures have been linked to a number of nucleic acid-related functions.

Guanine quadruplexes (G4) are a family of alternative nucleic acid structures, which have attracted the spotlights because of their stability under physiological conditions, and the widespread distribution of sequences compatible with G4 formation. The building brick of the guanine quadruplexes is a guanine *quartet*, a planar squared association of 4 guanines held by cyclic hydrogen bounds. The stacking of two or more quartets and the coordination of cations are responsible for the stability of the G4. The presence of runs of G is a requirement for G4 formation by a given nucleic acid sequence and will be the grounding of tools to identify quadruplex forming sequences (see later for a presentation of previous tools). The corpus of publications dealing with guanine quadruplexes is growing dramatically and many of these publications bring *in vivo* indications of quadruplex-related effects in telomere biology (2,3), transcription regulation (4), translation and RNA maturation (5,6), replication and genomic stability (7-9) and replication origine definition (10-13).

The interest for such structure and their genomic occurrence led to the development of several tools to predict quadruplex forming propensity starting with the seminal publications from the Balasubramanian and Neidle groups (14,15). The first generation of algorithms looked for pattern matching the stereotype $[G_nN_mG_nN_oG_nN_pG_n]$ expected to be favorable for quadruplex formation. In a second generation of algorithm, the group of Maizels proposed to look for the occurrence of runs of G_n ($n \geq 2$) in a window of a given size. Many variations have been proposed and applied to different types of genomic DNA or RNA databases. These algorithms are mainly looking for local enrichment of runs of G above a threshold size (n) of 2 or 3 usually (see (16) for review). Loop size has been (and is still) subject to discussion in the field: while the first studies constrained loop size between 1 and 7 nucleotides, latter developments allow longer loops, which was supported by experimental demonstration of formation of quadruplex with such long loops (up to 30 nucleotides in (17)).

In the meantime, a number of experimental studies established G4 formation for sequences that escape this consensus (false negatives; for example (18)) while a more limited number of articles reported sequences obeying the consensus and still unable to form G-quadruplexes *in vitro* (false positives) (17,19). Furthermore, many of these algorithms only provide a binary (yes / no; match / no match) answer which prevents any quantitative analysis that would have allowed putative correlation of a given quadruplex 'strength' metric with other genomic or functionnal parameters.

To overcome these limitations, we chose to develop a different algorithm, G4Hunter, that would take into account G-*richness* and G-*skewness* of a given sequence and give a score

(quadruplex propensity) as an output. To validate this model, we decided to benchmark this search using an unprecedented large number of sequences. We developed a combination of biophysical methods to accurately assess quadruplex formation *in vitro*. We validated this algorithm on the human mitochondria genome (16.6kb) because of its relatively high GC content and GC skewness as well as the biological relevance of QFS near instability hotpsots (20). We then applied the algorithm on a number of species including human, allowing us to conclude that the number of bona fide G4-prone sequences in the human genome must be very significantly reevaluated by a factor 2 to 10.

Material and methods:

Principle of the algorithm:

In order to take into account G richness and G skewness, each base is given a score between -4 to 4. The score is set to 0 for A and T (*i.e.*, neutral or indifferent), positive for G and negative for C. To account for G-richness (or C-richness, meaning G-richness on the complementary strand), a single G got a score of 1 while in a GG sequence, each G got a score of 2, in GGG sequence, each G got a score of 3, and in the sequence of 4 or more Gs, each G got a score of 4. The C get a similar scoring pattern, but with negative values. This allows to get a near-zero average score for G rich motifs in GC alternate sequences that are likely to form stable duplex that would compete with G4 formation. This design also allows to obtain the scoring for the complementary strand and/or for "C-quadruplexes", called i-DNA or i-motif, that are natural but less studied conterparts of G4. For a given sequence, the G4Hunter score (G4Hscore) is then the arithmetic mean of this "sequence" of numbers (see supplementary Figure S1A).

By construction, the G4Hscore should be centered arround 0 for random sequences, independently of GC content. This assumption is also verified on a number of genomes for which the sequence is not random. In contrast, the marked GC-*skewness* of the human mitochondrial genome leads to a non-nul average score, with a negative value of -0.4 as the light C-rich strand (L-strand) is reported in the databases.

Genomewide search:

For genomewide search, the mean of the scored nucleic acids sequence is then computed for a sliding window arbitrary set at 25 nucleotides. Regions in which this absolute value of this mean score rises above a threshold are extracted. The overlapping region are then fused and refined by removing non G (or non C) bases at each extremity, which could have passed through the windowing threshold procedure. The sequence may also be extended if the first or last base is a G (or a C). In that case, to avoid unwanted cut in G (or C)-run during the windowing threshold procedure, the previous and next bases, respectively, are also taken into account if necessary. The score of this fused and refined sequence is then computed again. This new score can end up below the threshold when for example two fused sequences shared a G-run rich core that at the end would have to compensate for two less favourable 'end' (see supplementary Figure S1B). G4FS density by kbp is then calculated by dividing the total number of G4FS by the length of the genome. To take into account gaps in the available genome release, bases that are N are not counted when evaluating genome size.

Scripting:

G4Hunter scores have been calculated either using Python or **R** scripts. Genomewide search have been performed usin **R** script taking advantages of exisiting packages such as GenomicRanges, rtracklayer, BSgenome, Biostrings and GenomicFeatures from BioConductor (21). Quadparser searches were also performed using home made version of the pattern searching script in **R** langage published in (16). ROC analysis was performed using the ROCR package for **R** (22).

Genome accession numbers and reference

G4FS search were performed on the human mitochondria genome (EF184640.1), 18 full genomes including human (hg19), mouse (mm10), fruitfly (dm3) and budding yeast (sacCer3) either from BSgenome packages for **R** (23) or from NCBI databases (see supplementary Table S4).

Compounds (oligonucleotides)
All oligonucleotides used in this research were purchased from Eurogentec (Seraing, Belgium) and stored at -20°C. Oligonucleotide strand concentrations were determined by absorbance at 260 nm using the extinction coefficients provided by the manufacturer. Sequences are provided in supplementary Table S2.

Thioflavin I fluorescent assay
Thioflavin T (3,6-Dimethyl-2-(4-dimethylaminophenyl) benzothiazolium cation) was purchased from Sigma-Aldrich (Ref. T3516) and used without further purification. Fluorescent assay was performed as described previously (24).

Nuclear magnetic resonance:
One-dimensional 1H NMR experiments were performed on a Bruker Avance 700 MHz spectrometer equipped with a liquid TXI 1H/13C/15N/2H, with Z- gradient probe, as described previously (25).

Circular dichroism (CD) spectroscopy:
Circular dichroism spectra were recorded on a Jasco J-815 equipped with a Peltier temperature control accessory (JASCO Co., Ltd., Hachioji, Japan) as described previously (25).

Absorbance spectroscopy (Tm, TDS and IDS):
All Spectra were recorded on a Uvikon XL spectrophotometer in a 10mM lithium cacodylate buffer (pH7.2) at 3 µM (except when stated otherwise) supplemented with KCl, NaCl or LiCl (100 mM, except when specified otherwise).

Thermal Difference Spectra (TDS) were obtained by taking the difference between the absorbance spectra from unfolded and folded oligonucleotides that were respectively recorded at high (>90°C) and low tempeature (4°C) (100 mM KCl). TDS provide specific signatures of different DNA structural conformations, provided that the structure is not too heat-stable (a number of G4 do not melt at high temperatures) (26).

Isothermal Difference Spectra (IDS) were obtained as described previously in (24) by taking the difference between the absorbance spectra from unfolded and folded oligonucleotides. These spectra were respectively recorded before and after potassium cation addition (100 mM KCl) at 20°C. IDS provide specific signatures of different DNA structural conformations.

UV melting experiments (Tm) were recorded as previously described (27,28). G4 unfolding is typically associated with a decrease in absorbance at 295 nm, giving an inverted transition at this wavelength.

Results:

Applying G4Hunter on a reference dataset extracted from literature

We performed a literature search of sequences for which quadruplex formation was experimentally confirmed or infirmed. We added our unpublished data and obtained a dataset of 392 sequences for which G4 formation propensy was properly tested. The sequences were separated into two groups (G4 and not G4) according to published results or our own previous testing. Sequence information is provided in supplementary Table S1 and the score using our algorithm (G4Hscore) is calculated for all the sequences. Quadparser analysis was performed in parallel using G-runs length setting of 2 or 3 and loops size of 7 or 12.

To first investigate the discriminative performance of the algorithm, we plotted the distribution of the G4Hscore separately for "G4" sequences and "not G4" sequences (Figure 1A). G4Hscore was significantly higher ($p<2\bullet10^{-16}$) for "G4" sequences than "not G4" sequences with average G4Hscore values of 1.64 ± 0.46 and 0.16 ± 0.66 respectively. The significance of the observed distribution differences was performed using the non-parametric Wilcoxon rank-sum / Mann-Whitney U-test (null hypothesis: distributions are not different). From the histogram of the score distribution for this reference dataset (Figure 1B), a threshold of 1 allows a good discrimination of G4 *vs*. notG4 sequences.

In order to evaluate the quality of our scoring system, we performed a Receiver Operating Characteristic (ROC) analysis. The ROC curve is characteristic of a good estimator with prononced convexity toward true positive results. Random performing estimator would follow the dotted diagonal. The area under the ROC curve (AUC) estimates the accuracy of the algorithm to discriminate between G4 and notG4 sequences to more than 0.96 (an area of 0.5 represents a random, useless and non discriminating value whereas an AUC of 1 indicates a perfect prediction). To estimate the threshold "quality", we indicated the ROC result for 5 different threshold between 1 and 2 and also the position of the ROC analysis for 3 settings commonly used for QuadParser. The threshold value of 1 for G4Hunter provides both highest sensitivity and the highest sensibility (Figure 1C). All values of threshold performed better than Quadparser with G-runs of 2, while Quadparser with G-runs of 3 seems to have a lesser false positive rate. One has to take into account that pattern type motifs have been long though to hold for true G4 motif, and thus the reference dataset is largely biased toward this type of sequences. Nevertheless, by accepting a slightly higher rate of false positive (up to 10%), threshold of 1 or 1.2 for G4Hunter allows the recovery of more "true" G4FS.

We next evaluate the precision, *i.e.* the fraction of sequence forming a G4 detected experimentally among the sequences found by Quadparser or with a G4Hscore above a threshold (Figure 1D). This analysis also reveals a huge bias in our "reference" library toward G4FS (76%) and toward G4FS fitting with the "classical" definition of quadruplexes forming sequence (50%): sequences with at least 4 runs of at least 3 Gs separated by loop up to 7 bases *i.e.*, the classical Quadparser parameters.

Applying G4Hunter on the human mitochondria genome and validation

In order to strengthen the analysis, we decide to use G4Hunter to identify G4FS in a less biased (or more accurately, differentially biased) dataset in which we could perform an exhaustive search for G4 propensity. We chose the human mitochondrial genome (release EF184640.1 from NCBI) because it is relatively compact (16.6 kbp circular DNA molecule), GC rich (and therefore potentially more prone to G4 formation). It was also the first

eukaryotic genome to be sequenced (29); it encodes for necessary subunits of mitochondrial subunits I, II, III, IV and V. The genome consists of a heavy (H) strand, rich in guanine bases, and a complementary light (L) strand, which is rich in cytosines. The first G4 motifs analysis was performed in yeast mitochondria (30). Several studies found a correlation between transcription termination, primer formation and R-loop stability and the presence of G4 structures (31). Computational analyses correlate mtDNA deletions with non-B DNA propensity (32). Mitochondrial DNA deletions are prominent in human genetic disorders and Quadparser-predicted mitochondrial G-quadruplex-forming sequences map in close proximity to known deletion breakpoints (20,33). A striking feature of the human mitochondrial genome is its GC skewness; this is reflected by the average G4Hscore value (-0.4) and the distribution of scores (see next §).

To identify G4FS with G4Hunter, we choose to first calculate G4Hscores as means on a sliding window of 25 nt. We choose this window size as it is a close match to the mean size (26 nucleotides) of the G4FS reference dataset. Genomic search was then performed according to the procedure reported in materials and methods. According to the results from the analysis of the reference dataset, a threshold of 1 was chosen. It means that we selected every sequence for which the G4Hscore was above 1 in absolute value; to evaluate G4FS on both strand simultaneously. 1846 overlapping sequences of 25 nucleotides were identified, corresponding to 169 windows of consecutive values above the threshold. Because of the mitochondrial genome GC skewness only 4 (2.3%) are found on the L strand (Ghscore ≥ 1), in contrast with 166 located on the H strand (G4Hscore ≤ -1). Out of these sequences, 165 were suitable for biophysical evaluation (see supplementary Table S2).

In parallel, we applied Quadparser to identify G4FS on the same human mitochondrial genome with the following settings: G-runs of 2 or more and loops up to 7 nt (QP27), G-runs of 3 or more and loop up to 7 nt (QP37) and G-runs of 3 or more and loop up to 12 nt (QP312). These parameters allowed us to identify 81, 5 and 11 candidates, respectively. All the hits founds by Quadparser with G-runs of 3 and more were also found G4Hunter. More than fifty sequences were found by G4Hunter and Quarparser with G-runs of 2 and more (Figure 2A). The sequences found by Quadparser but not by G4Hunter are the sequences characterized by a score below 1 and/or length less than 25 nucleotides. The 26 sequences in this category are listed in supplementary Table S2. For experimental validation, we also added 7 sequences of 25 nt with a score between 0 and 0.4 that are expected to behave as negative controls.

We performed the biophysical characterization to identify *bona fide* G4 forming sequences in this dataset of 198 sequences. Our initial misconception was that a limited number of rapid assays would be sufficient for confirm G4 formation for a given sequence. This turned out to be far more difficult than anticipated. In order to obtain a reliable answer to the seemingly simple question *"is this oligonucleotide forming a quadruplex in vitro?"* we had to combine the results of 5-6 different techniques (IDS, TDS, UV-melting, CD, NMR and Thioflavin assay). Taken individually, none of methods gave a satisfactory answer in all cases (IDS and 1D-NMR being the most reliable techniques). The experimental results for all sequence tested are presented in supplementary information ("Supplementary results" for close to 200 different oligonucleotides): each page refers to a single sequence, and all assays performed on it are presented in a similar fashion for each sequence. The conclusion is indicated for each oligonucleotide.

Based on the conclusion from these tests, these sequences were classified as G4 (n=128) and not G4 (n=70) and subject to the same statistical analysis as the reference dataset

(see supplementary informations and supplementary Figure S2A-C). The discriminating score appears to be 1.2 (Figure S2D) but G4Hscore are less dispersed as the one from the reference dataset as most of the sequences have been selected with a threshold of 1.

This study allows us to conclude that on a real dataset, G4Hunter score is a good estimator of the ability of a sequence to fold into a quadruplex. To evaluate the ability of G4Hunter to identify *bona fide* G4FS, we adjusted the G4Hunter list result to avoid redundance by merging overlapping sequences (if they are on the same strand) using thresholds ranging from 1 to 2. We obtained lists of 96, 67, 25, 19 and 7 hits, respectively, for threshold values of 1, 1.2, 1.5, 1.7 and 2. To evaluate the ability to form G4, we looked for overlaps between these sequences and the list of 128 sequences characterized biophysically to form G4. These 128 characterized G4 represent a fairly complete picture of the G4 on the mitochondria genome. This allows to compare the precision of the different algorithms and settings (Figure 2B). G4Hunter with a threshold of 1.2 seems to be the best compromise in number of hit with a good precision and thus a low false discovery rate (FDR = 9%).

We then decided to study to impact of window size on G4Hunter performance. To that aim, we computed the G4Hscore using running windows of variable length, from 15 to 100 nucleotides. The lower limit of 15 nucleotides corresponds to the length of one of the shortest known stable intramolecular quadruplex (TBA; the thrombin binding aptamer, d-GGTTGGTGTGGTGG). Most of the intramolecular quadruplexes are much shorter than 100 nucleotides; the 100 nt upper value was chosen to match the window size of 100 nucleotides chosen by Maizels and colleagues (34). When evaluating the number of potentiel G4FS in the mitochondrial genome for various window sizes and thresholds, the highest number of candidates are obtained with the couple window/threshold of 15/1 but with a low precision (50%) (see supplementary Table S3). Increasing window size reduces the number of G4 forming sequences. On the other hand increasing window size and threshold increases the precision of the detection, but to the cost of a reduction of the number of sequences identified. If we set a precision threshold of 90%, maximum of G4 are found with the settings 25/1.2 and 15/1.6. If this threshold is set at 95%, the best settings are 15/1.7.

For further studies, with the aim to maximize the chance of identifying G4-forming sequences, we chose to use a window size of 25 nt with thresholds values between 1 and 2 (a higher threshold means the sequence is very likely to form a quadruplex, but at a cost of more false negatives).

Analysis of whole genomes

Using G4Hunter we identified G4FS with different threshold for 20 different genomes including the classical model organisms *Homo sapiens*, *Mus musculus*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Plasmodium falciparum*, *Escherichia coli* and *Arabidopsis thaliana*. We also included *Dictyostelium discoideum* as a GC poor (22%) organism.

The number of potentiel G4FS is of course dependent on genome size (supplementary Table S5A-B). For this reason, to allow a meaningful comparison between various species, we computed the density of G4FS per kb. The occurence of G4FS is comparable and the highest among mammalians, ranking from 2.5 ± 0.1/kb with a threshold of 1 (likely G4FS) to 0.17 ± 0.04/kb with a threshold of 2 (very stable/highly likely G4FS) (Table 1A-B). The relationship between G4FS density and threshold can be described with an exponential fit (Figure 3) with a similar exponential factor (slope) of -2.71 ± 0.25 for mammalians.

Unsurprisingly, the human, chimpanzee and macaque genomes have very close properties regarding the number of G4FS, their density and the exponential factor for the dependency toward the threshold (-3.00 ± 0.07). Outside the mamalian class, the densities of G4FS and their dependency to the threshold are quite diverse. Four families of curves can be proposed (Figure 3):

i) G4FS rich genomes (mammals, rice, chicken, supplementary Figure S3A);

ii) intermediate G4FS genome (*E. coli*, fly, bee and zebrafish, supplementary Figure S3B);

iii) low G4FS genomes (budding and fission yeasts, *C. elegans* and *A. thaliana*, supplementary Figure S3C) and

iv) very poor G4FS genomes (*P. falciparum* and *D. discoideum*, supplementary Figure S3D).

For the first three classes, an exponential fit gives a very good description of the density *vs*. threshold distribution. Interestingly, the slope varies between -2.5 (*D. melanogaster* (-3), *A. mellifera* (-2.4)) and -5 (*E. coli* (-5.6), yeasts (-5.2,-5.1) and *A. thaliana* (-5.1)). The GC poor organism *Dictyostelium discoideum* and *Plasmodium falciparum* exhibit the lowest density of G4FS. For these two genomes an exponential fit fails to describe perfectly the whole density *vs*. threshold curve with a marked deviation for threshold above 1.4 (supplementary Figure S3E). Above this limit, the G4FS density for *D. discoideum* decreases less markely whereas the one of P. falciparum decreases more. This could indicate a conservation of G4FS with high G4Hscore in *D. discoideum* arguing for a biological function of these type of sequence in this organism. On the other hand, the marked decrease for *P. falciparum* for high thresholds as well as the slopes for *E. coli*, both yeasts and *A. thaliana* could indicate a counterselection against very stable quadruplexes in these organisms.

Results can be easily visualized with a genome browser like IGV (35) both at a promoter scale (Figure 4, top and supplementary Figure S4A) or larger genomic scale (Figure 4, bottom and supplementary Figure S4B). In the *MYC* case (Figure 4, top), we can compare the G4FS found by different settings of G4Hunter and Quadparser. Candidate sequences are found with G4Hunter (with a threshold of 1.5 or more), that were not found with the classical Quadparser parameters (QP37). At a larger scale (Figure 4, bottom), a local enrichment in G4FS near *MYC* is striking, as well as near *MIR1204*. The profiles shown in supplementary information confirm that some genomic regions, such as the loci near *HRAS* or *SRC* (supplementary Figure S4B) are richer in G4FS. Regarding the budding yeast genome, the snapshots presented in supplementary Figure S4C, left illustrate the paucity in G4FS and allowed us to identify the G4FS characterized in (30). On the fruitfly genome (supplementary Figure S4C, top-right), we illustrate a local enrichment in G4FS near the heterochromatin region recognized by a G4 specific antibody (36). The bottom right part of the supplementary Figure S4C illustrates the local enrichment of G4FS in the rDNA cluster of the human genome (37).

Distribution in the human genome

In order to get a better resolution for G4FS in the human genome, we performed a search using three window sizes (15, 20 and 25 nucleotides) which gave a good compromise between number of hits and precision for the mitochondria genome. In all three cases, the number decrease exponentially with the threshold with a similar slope (supplementary Figure S4A, hits number *vs.* threshold). Based on the mitochondria results, with a window size of 25 and a threshold of 1.2, the precision is above 90%, meaning that more that 90% of the sequences found with theses setting should form a quadruplex. We compared the results

found with different window sizes but leading to similar number of hits (around 4 millions): 25/1.2; 20/1.36 and 15/1.64 (see supplementary Figure S4B). The majority of the hits are found by the three settings (respectively 78%, 77% and 69%). Given that most G4-forming sequences described in the literature are longer than 20 nucleotides, we decided to focus the next analysis on the results obtained with a window of 25.

To analyze the repartition of G4FS within the genome using genomic features such as promoter, CDS, introns or exons we decide to compute three metrics: *i)* the fraction of the feature containing one or more G4FS: *ii)* the fraction of the G4FS found in a genomic feature and finally *iii)* the local density of G4FS per kb in the genomic features.

In order to get a background distribution to test the enrichment of G4FS in a given genomic feature, we used the following strategies:

- for metrics *ii)* and *iii)*, we computed the total coverage of the feature on the genome by dividing the sum of the length of the feature by the size of the genome. At this point it should be noted that all the genome is not sequenced in the hg19 release, leaving large domains of unknown sequence, for which sequence-based search such as G4Hunter (and also Quadparser) could not find any hit. For example, the first 16 and 19Mb of chromosomes 22 and 14 are all Ns. A similar 32Mb gap is found on the Y chromosome. To evaluate enrichment, we thus adjusted the size of the genome to exclude all N (ambiguous bases). We also excluded the mitochondrial DNA from this analysis as its transcripts are not annotated in the UCSC Known Genes list. This led us to reduce the known human genome size from $3.1 \cdot 10^9$ to $2.86 \cdot 10^9$ bases. We called this background the global background.

- for metrics *i)*, *ii)* and *iii)*, in order to remain closer to the real genome composition, G4FS lists were randomized 1000 times in the available and annotated genome using home made R script (we took care to avoid the genomics gaps filled only with N if the size of those gaps was above 20). This allowed us to generate a family of 1000 lists with the same size and following the same strand, width and chromosomal distribution for each G4FS list. We called this background the resampled background.

Promoters

Using genomic annotation from the known genes list of UCSC, we defined the promoter region as the 1kb before the TSS of the unambiguous transcripts. We obtained 39692 unique promoters.

i) When looking at the occurrence of G4FS in promoters, choosing a low G4-Hunter threshold (1 or 1.2) or low stringency Quadparser settings (QP27), led to the conclusion that almost all the promoters contains at least one G4FS (94, 84 and 94% respectively). However, these results are close to the fraction obtained for the mean of resampled background (93, 77 and 96% respectively). On the other hand, when using more stringent criteria (G4H with a threshold of 1.5, 1.75 and 2), the fractions of promoters with at least one G4FS are 66, 52 and 37% respectively. These figures are significantly higher than the fractions obtained from the randomized lists (45, 27 and 14% respectively) and correspond to enrichments above the randomized background of 1.5, 2 and 2.6-fold, respectively. Similar results are obtained with QP37 and QP312 (see Table 2). Thus, the notion that promoters are enriched in G4FS only holds true when considering "stable" quadruplexes only, those found either by G4Hunter with a threshold above 1.5 or by Quadparser with blocs of at least 3 guanines (QP312 and QP37).

ii and iii) Promoters represent thus 1.4% of the genome but as promoters can overlap, this figure can be reduced to 1.2% of the total genome. If G4FS repartition were random, one would expect that 1.2% of the G4FS would be found in promoters. It is neither the case for G4FS list extracted by G4Hunter with threshold from 1 to 2 or by Quadparser: a 2.2 to 5.5-fold enrichment is found in all cases (Table 2). We also computed the occurrence of the hits in promoters for the resampled background for each list. This allows us to conclude that the fraction of G4FS found in promoters by G4Hunter or Quadparser is 2 to 4.5-fold higher than the fraction of random sequences following the the same chromosomal, length and strand distribution with a p-value $< 10^{-3}$ (p-values were obtained from a null distribution based on resampled background). A similar conclusion is reached when studying quadruplex density (number of G4FS per kb, see table Table 2) using either the global density of G4 in the genome (enrichment between 2.2 and 5.5) or the density computed by resampling (enrichment between 2 and 4.6).

This promoter enrichment, which is in agreement with previously published studies, also reflects the bias toward G/C or CpG island enrichment in promoters. Using the strand information and the position, we could generate the local profile of all G4FS arround TSS (±1kb) (Figures 5 and S5C). A clear local enrichment between -200 bp and 0 before the TSS is observed on both strands of the promoter. However, the most striking feature is the asymetry between the coding and not coding strands in the first 500bp after the promoter with an enrichement on the coding strand. This could reflect enrichment either in a post TSS part of the promoter or in the 5' UTR of the transcripts.

Other transcription-related genomic features

G4FS are enriched in promoters. As briefly discussed before, this may be the consequence of a transcriptional role for G4 structures, or be an indirect consequence of the typical properties of promoters such as the G/C richness, the presence of CpG islands, and/or the presence of G/C rich transcription factor consensus binding sites. We thus performed the same analysis for CDS, 5'UTR, 3'UTR, exons (unique, first, other and last), intron (unique, first, other and last), and 100bp around each exon/intron and intron/exon junctions (for unique, first, other and last introns). An obvious enrichment in G4FS is observed in the 5'UTR, first exons and first exon/intron junctions (see supplementary Table S6-8). This confirms the observation from the distribution profile around the TSS and point out also to a potential role of G4FS at the first splicing junction (38). Profile of such junction is represented on Figure 6 (and S6) and indicates a local enrichment in G4FS on the coding strand on the intronic part of the junction.

Discussion:

The constant increase in interest for Quadruplex structures, together with the available genome-wide and transcriptome wide methods highlights the need for accurate tools to predict G-quadruplex propensity. While several algorithms are currently available, a number of studies have established G4 formation for sequences undected by these tools (false negatives) while a more limited number of examples reported false positives. These observations prompted us to propose a new algorithm, based on G/C skewness and the presence of G-blocks. Rather than defining an arbitrary limit on the number of consecutive guanines required or on loop size / window size, we chose to define a scoring function that reflects quadruplex propensity.

In contrast with many other previous studies, we experimentally validated this work on large set of sequences. This validation step turned out to be far more time consuming than expected: reaching a firm conclusion on the hundreds of "real life" sequences investigated here appeared more challenging than when we were working on "well-behaved" sets of oligonucleotides, for example when containing exactly four blocs of 3 guanines. We had to rely on a set of no less than 6 independent methods to reach an unambiguous conclusion for 95%+ of the sequence tested. This combination of techniques is a proposed standard for experimental validation, but can be also completed by other methods.

While Quadparser gives a "yes" or "no" answer, G4Hunter gives a score. The user may choose different threshold values to optimize the search. A high threshold (1.5 or more) will minimize the number of false positive and favor highly stable quadruplex motifs, but will ignore a number of true G4-forming sequences. To obtain a more exhaustive count of G4 potential, a lower threshold is recommended. 1.2 seems a good compromise, but will still miss true G4 motifs. This parameter can be easily adjusted and will also depend on the application: for example, one could imagine applications (DNA origami, PCR) in which multiple oligonucleotide staples or primers are used: imposing a G4HS value below 1.5 (or, to be on the safe side, 1.2) should minimize artefacts due to undesired G4 formation. On the other hand, when a structural analysis of a sequence is required (for example, after SELEX) a G4HScore above 0.9 should prompt to test the G4 hypothesis. A number of aptamers are known to adopt a quadruplex conformation, and our unpublished results (Kuznetsov *et al*, in preparation) demonstrate that the real number of aptamers found in the litterature which do form quadruplexes is even higher. Whatever the chosen threshold, one should remember that the discrimination is not perfect, and that a few outliers will be treated in the wrong category. Figure 1B illustrates that this problem is relatively limited.

Another important parameter is window size. Most of the analyses presented here were performed with a default size of 25 nucleotides. This more or less corresponds to the actual size of many experimentally determined quadruplexes. Nevertheless, this parameter can easily be adjusted, down to 15-20 nucleotides and with well defined upper limit. Using large windows of 100+ nucleotides may actually help identifying genomic regions in which multiple contiguous G4 can be formed, rather than individual structures. This may well be relevant for some biological effects: Rodriguez *et al* found an increased number of γH2AX foci in regions where multiple Quadparser hits are found (39).

Regarding genome-wide studies, intramolecular structures are probably more biologically relevant. In contrast, for all DNA *in vitro* applications such as the one mentioned above (PCR, Origami, Selex) all types of molecularities should be considered: G4 formation may hamper

PCR independently of whether the quadruplex is composed of 1, 2 or 4 independent strands. While G4Hunter will mostly be used to identify intramolecular structures, the current search parameters and the experimental validation do not explicitly exclude species of higher molecularities (dimers, etc…). Search parameters can be slightly altered to favor intramolecular forms, for example by imposing a minimum number of 8 guanines within the window (with window sizes of 20 and 25 nucleotides, a G7 tract embedded in a A/T rich region would give a score of 28/20 or 28/25 = 1.4 or 1.12, respectively and therefore be selected as a G4-prone motif, even if obviously unable to form an intramolecular quadruplex). This example also illustrates that short window sizes tend to make this problem more acute: the shorter the window, the more blatant the contribution of a single long G-run.

Bearing in mind the possibility to adjust these two key parameters, we can compare the performance of G4Hunter with 3 different versions of Quadparser (QP37, QP312 or QP27). Interestingly, while stringent versions of Quadparser gave an excellent false positive rate (0 in the mitochondria data set: all sequences predicted experimentally form G4), they miss a large majority of true quadruplex forming motifs. QP27 finds a lots more, at the cost of accuracy (80%, Figure 2B). Interestingly, G4H1.2 has a much higher precision (91%) with a comparable number of hits. In other words, the number of false positives is more than halved with G4H1.2 as compared to QP27.

Provided a reasonable G4HScore threshold is chosen, the number of hits found at the genome wide level is higher than QP37. The number of G4-prone sequences found in the human genome should therefore be significantly reevaluated as compared to the commonly accepted figure of 376,000 G4FS. Remarkably, Balasubramanian and colleagues came up with a higher figure using the G4seq approach (40). To obtain a nearly identical number of sequences with G4Hunter, one has to select a threshold value of 1.75 with a window size of 25 (1.96 if the window size is reduced to 20). It is interesting to compare this figure with the number of sequences predicted to be genomically unstable (18153) according to Nicolas and colleagues (41): these two values suggest that a large majority of quadruplex-prone motifs are genomically stable, arguing that one can imagine regulatory functions for such sequences without a concomittant deleterious instability.

While capable of finding far move G4 motifs without sacrifying the false positive rate, statistical analysis demonstrates that G4Hunter is not perfect. Initial values chosen for Gn blocks were the corresponding integer values. This facilitates calculations when large genomes are considered. However, there is no reason not to consider fractional values for these parameters, and we are currently investigating this possibility. Having a relatively large database of sequences of our own, we should be able to recalculate accuracy in each case. Furthermore, the values attributed for the other bases (A, T, C) are context independent; this is obviously an approximation: for example a GGGCGGGN…NGGGCGGG can form a very stable quadruplex, and the two 1-nt long cytosine loops are actually more favorable than single thymines or adenines (42): giving a negative score for C and null for A and T is not justified for this family of sequences. We therefore hope that our work will stimulate further studies aimed at improved G4Hunter!

Acknowledgments:

We are grateful to the genotoul bioinformatics platform Toulouse Midi-Pyrenees for computing and storage resources. L.L. wishes to thank all the Cuvier team, in particular P.G.P. Martin, G. Micas, Y. Visser and W.T. Jones for support using **R**.

Funding:

This work was supported by Fondation ARC (subvention libre to J.L.M.), Conseil régional d'Aquitaine (fellowship awarded to A.B.), *Agence Nationale de la Recherche* ("OligoSwitch" [ANR- 12-IS07–0001], "Quarpdiem" [ANR-12-BSV8–0008–01], and "VIBBnano" [ANR-10-NANO-04–03] to J.L.M.) and Université de Toulouse ("PISAAN" [AO1-2013] to L.L.).

References:

1. Watson, J.D. and Crick, F.H. (1953) Genetical implications of the structure of deoxyribonucleic acid. *Nature*, **171**, 964-967.
2. De Cian, A., Lacroix, L., Douarre, C., Temime-Smaali, N., Trentesaux, C., Riou, J.F. and Mergny, J.L. (2008) Targeting telomeres and telomerase. *Biochimie*, **90**, 131-155.
3. Zimmermann, M., Kibe, T., Kabir, S. and de Lange, T. (2014) TRF1 negotiates TTAGGG repeat-associated replication problems by recruiting the BLM helicase and the TPP1/POT1 repressor of ATR signaling. *Genes Dev*, **28**, 2477-2491.
4. 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. *PNAS*, **99**, 11593-11598.
5. Wieland, M. and Hartig, J.S. (2007) RNA quadruplex-based modulation of gene expression. *Chem Biol*, **14**, 757-763.
6. Millevoi, S., Moine, H. and Vagner, S. (2012) G-quadruplexes in RNA biology. *Wiley interdisciplinary reviews. RNA*, **3**, 495-507.
7. Cheung, I., Schertzer, M., Rose, A. and Lansdorp, P.M. (2002) Disruption of dog-1 in Caenorhabditis elegans triggers deletions uptstream of guanine-rich DNA. *Nat. Genet.*, **31**, 405-409.
8. Lopes, J., Piazza, A., Bermejo, R., Kriegsman, B., Colosio, A., Teulade-Fichou, M.P., Foiani, M. and Nicolas, A. (2011) G-quadruplex-induced instability during leading-strand replication. *EMBO J*, **30**, 4033-4046.
9. 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.
10. Besnard, E., Babled, A., Lapasset, L., Milhavet, O., Parrinello, H., Dantec, C., Marin, J.M. and Lemaitre, J.M. (2012) Unraveling cell type-specific and reprogrammable human replication origin signatures associated with G-quadruplex consensus motifs. *Nat Struct Mol Biol*, **19**, 837-844.
11. Cayrou, C., Coulombe, P., Puy, A., Rialle, S., Kaplan, N., Segal, E. and Mechali, M. (2012) New insights into replication origin characteristics in metazoans. *Cell Cycle*, **11**, 658-667.
12. 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.
13. Comoglio, F., Schlumpf, T., Schmid, V., Rohs, R., Beisel, C. and Paro, R. (2015) High-resolution profiling of Drosophila replication start sites reveals a DNA shape and chromatin signature of metazoan origins. *Cell reports*, **11**, 821-834.
14. Huppert, J.L. and Balasubramanian, S. (2005) Prevalence of quadruplexes in the human genome. *Nucleic Acids Res*, **33**, 2908-2916.
15. Todd, A.K., Johnston, M. and Neidle, S. (2005) Highly prevalent putative quadruplex sequence motifs in human DNA. *Nucleic Acids Research*, **33**, 2901-2907.
16. Hershman, S.G., Chen, Q., Lee, J.Y., Kozak, M.L., Yue, P., Wang, L.S. and Johnson, F.B. (2008) Genomic distribution and functional analyses of potential G-quadruplex-forming sequences in Saccharomyces cerevisiae. *Nucleic Acids Research*, **36**, 144-156.
17. Guedin, A., Gros, J., Alberti, P. and Mergny, J.L. (2010) How long is too long? Effects of loop size on G-quadruplex stability. *Nucleic Acids Res*, **38**, 7858-7868.
18. Mukundan, V.T. and Phan, A.T. (2013) Bulges in G-quadruplexes: broadening the definition of G-quadruplex-forming sequences. *J Am Chem Soc*, **135**, 5017-5028.

19. Guedin, A., Alberti, P. and Mergny, J.L. (2009) Stability of intramolecular quadruplexes: sequence effects in the central loop. *Nucleic Acids Res*, **37**, 5559-5567.
20. Bharti, S.K., Sommers, J.A., Zhou, J., Kaplan, D.L., Spelbrink, J.N., Mergny, J.L. and Brosh, R.M., Jr. (2014) 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*, **289**, 29975-29993.
21. Huber, W., Carey, V.J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B.S., Bravo, H.C., Davis, S., Gatto, L., Girke, T. *et al.* (2015) Orchestrating high-throughput genomic analysis with Bioconductor. *Nat Methods*, **12**, 115-121.
22. Sing, T., Sander, O., Beerenwinkel, N. and Lengauer, T. (2005) ROCR: visualizing classifier performance in R. *Bioinformatics*, **21**, 3940-3941.
23. Pages, H. R package version 1.36.3 ed.
24. Renaud de la Faverie, A., Guedin, 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.
25. Amrane, S., Kerkour, A., Bedrat, A., Vialet, B., Andreola, M.L. and Mergny, J.L. (2014) 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*, **136**, 5249-5252.
26. 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.
27. Mergny, J.L., Phan, A.T. and Lacroix, L. (1998) Following G-quartet formation by UV-spectroscopy. *FEBS Letters*, **435**, 74-78.
28. Mergny, J.L. and Lacroix, L. (2009) UV Melting of G-Quadruplexes. *Curr Protoc Nucleic Acid Chem*, **Chapter 17**, Unit 17 11.
29. Anderson, S., Bankier, A.T., Barrell, B.G., de Bruijn, M.H., Coulson, A.R., Drouin, J., Eperon, I.C., Nierlich, D.P., Roe, B.A., Sanger, F. *et al.* (1981) Sequence and organization of the human mitochondrial genome. *Nature*, **290**, 457-465.
30. Capra, J.A., Paeschke, K., Singh, M. and Zakian, V.A. (2010) G-quadruplex DNA sequences are evolutionarily conserved and associated with distinct genomic features in Saccharomyces cerevisiae. *PLoS computational biology*, **6**, e1000861.
31. Wanrooij, P.H., Uhler, J.P., Simonsson, T., Falkenberg, M. and Gustafsson, C.M. (2010) G-quadruplex structures in RNA stimulate mitochondrial transcription termination and primer formation. *PNAS*, **107**, 16072-16077.
32. Damas, J., Carneiro, J., Goncalves, J., Stewart, J.B., Samuels, D.C., Amorim, A. and Pereira, F. (2012) Mitochondrial DNA deletions are associated with non-B DNA conformations. *Nucleic Acids Res*, **40**, 7606-7621.
33. Dong, D.W., Pereira, F., Barrett, S.P., Kolesar, J.E., Cao, K., Damas, J., Yatsunyk, L.A., Johnson, F.B. and Kaufman, B.A. (2014) Association of G-quadruplex forming sequences with human mtDNA deletion breakpoints. *BMC Genomics*, **15**, 677.
34. Eddy, J. and Maizels, N. (2006) Gene function correlates with potential for G4 DNA formation in the human genome. *Nucleic Acids Res*, **34**, 3887-3896.
35. Robinson, J.T., Thorvaldsdottir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G. and Mesirov, J.P. (2011) Integrative genomics viewer. *Nat Biotechnol*, **29**, 24-26.
36. Hoffmann, R.F., Moshkin, Y.M., Mouton, S., Grzeschik, N.A., Kalicharan, R.D., Kuipers, J., Wolters, A.H., Nishida, K., Romashchenko, A.V., Postberg, J. *et al.* (2015) Guanine quadruplex structures localize to heterochromatin. *Nucleic Acids Res*.

37. Drygin, D., Siddiqui-Jain, A., O'Brien, S., Schwaebe, M., Lin, A., Bliesath, J., Ho, C.B., Proffitt, C., Trent, K., Whitten, J.P. *et al.* (2009) Anticancer activity of CX-3543: a direct inhibitor of rRNA biogenesis. *Cancer Res*, **69**, 7653-7661.
38. Eddy, J. and Maizels, N. (2008) Conserved elements with potential to form polymorphic G-quadruplex structures in the first intron of human genes. *Nucleic Acids Res*, **36**, 1321-1333.
39. Rodriguez, R., Miller, K.M., Forment, J.V., Bradshaw, C.R., Nikan, M., Britton, S., Oelschlaegel, T., Xhemalce, B., Balasubramanian, S. and Jackson, S.P. (2012) Small-molecule-induced DNA damage identifies alternative DNA structures in human genes. *Nat Chem Biol*, **8**, 301-310.
40. Chambers, V.S., Marsico, G., Boutell, J.M., Di Antonio, M., Smith, G.P. and Balasubramanian, S. (2015) High-throughput sequencing of DNA G-quadruplex structures in the human genome. *Nat Biotechnol*, **33**, 877-881.
41. Piazza, A., Adrian, M., Samazan, F., Heddi, B., Hamon, F., Serero, A., Lopes, J., Teulade-Fichou, M.P., Phan, A.T. and Nicolas, A. (2015) Short loop length and high thermal stability determine genomic instability induced by G-quadruplex-forming minisatellites. *EMBO J*, **34**, 1718-1734.
42. Guédin, A., De Cian, A., Gros, J., Lacroix, L. and Mergny, J.L. (2008) Sequence effects in single-base loops for quadruplexes. *Biochimie*, **90**, 686-696.

Figure Legends:

Figure 1:
A- Boxplot of the G4Hscore for the reference dataset. Open circles represent the G4Hscore values for individual sequences belonging to either G4 or not G4 classes. P-value for a Wilcoxon test between the 2 classes is indicated. **B-** Histogram of density distribution of the absolute values of the G4Hscore for the two classes of the reference dataset (Blue: G4 forming class, Red: not G4 forming class). The green dotted line indicates the value of abs(G4Hscore) for which more G4 than not G4 are found in this density histogram. **C-** ROC curve for G4Hunter on the reference dataset. Black symbols represent the position of individual threshold values for G4Hunter. Green, blue and red crosses represent position of the corresponding ROC values after applying QuadParser algorithm on the reference dataset with the following settings: runs of 2Gs and loops length between 1 and 7 (QP27, green), runs of 3Gs and loops length between 1 and 7 (QP37, blue) and runs of 3Gs and loops length between 1 and 12 (QP312, red). **D-** Precision vs threshold for G4Hunter. Fraction of sequence classified as G4 forming and which abs(G4Hscore) is above the threshold in X-axis. Precision for the threshold 1, 1.2 and 1.5 are indicated with dotted line in purple, orange and black respectively. Precision with QP27, QP37 and QP312 are indicated in green, blue and red respectively.

Figure 2:
A- Euler diagram representation of sequence from the human mitochondrial genome found by G4Hunter with a threshold of 1 (G4H1, blue), sequences found by quadparser (runs of 2Gs and loops length between 1 and 7, QP27, red) and sequences experimentally demonstrated to form a G4 (green). Numbers indicate population of each subclass. **B-** Number of sequences found by the different algorithms and setting in the mitochondria genoma (G4, in blue, not G4, in red). The percentages in white in the blue bar indicate the fraction of Hits for which G4 formation was experimentally confirmed. Note that in this figure, the number of sequences for each list is the number of non-overlapping sequences.

Figure 3:
Global G4FS density (number of G4Hunter hits by kb) *vs* threshold for whole genomes. The number of hits found by G4Hunter using a window of 25 was computed at different thresholds from 1 to 2 for whole genomes of *Homo sapiens* (hg19, blue), *Drosophila melanogaster* (dm3, red), *Saccharomyces cerevisiae* (SacCer3, pink) and *Dictyostelium discoideum* (ddAX4, green). The density of hits by kb is represented with respect to the threshold used and was fitted using an exponential fit. The fitted equation are represented with the same color as the genome.

Figure 4:
Genome browser view of the G4FS found by different algorithm near the *MYC* promoter. G4FS on a 4kb region (top) and G4FS density (#G4FS/kb) on a 200kb region (bottom) are represented for G4Hunter with the threshold of 1.2 (pink), 1.5 (green), 1.75 (dark blue) and 2 (light blue). G4FS from QP27, QP312 and QP37 are represented in grey, orange and red respectively.

Figure 5:
Profiles of G4FS around the transcription start site (TSS) for the UCSC Known Genes list with 4 different thresholds for G4Hunter (1.2; 1.5; 1.75 and 2 for the upper left, upper right, lower left and lower right cadrans, respectively). The number on the Y axis, the fraction of

G4FS, represents at the nucleotide level for each position the number of times this nucleotide is found in a G4FS divided by the number of TSS region (39692). The blue and red curve correspond to the G4FS found on the non-coding and coding strands, respectively.

Figure 6:
Profiles of G4FS arround the first exon/intron junction for transcript of the UCSC Known Genes list with 4 different threshold of G4Hunter (1.2; 1.5; 1.75 and 2 for the upper left, upper right, lower left and lower right cadrans, respectively). The number on the Y-axis, the fraction of G4FS, represents at the nucleotide level for each position the number of time this nucleotide is found in a G4FS divided by the number of juction region (37466). The blue and red curves correspond to the G4FS found on the non-coding and coding strands, respectively.

Tables:

Table 1: Number of hits per kbp of the sequenced genome obtained with G4hunter using a window of 25 nucleotide and a threshold indicated in the first column. Note that to calculate the length of the sequenced genome, unattributed bases N have been excluded.

Threshold	*H.sapiens*	*M. mus.*	*D. mel.*	*C. elegans*	*D. discoi.*	*S. cer.*	*S. pombe*	*P. falc.*	*E. coli*	*A. thaliana*
1	2.425	2.329	1.575	0.817	0.289	0.698	0.544	0.178	1.301	0.704
1.25	1.010	1.027	0.609	0.268	0.078	0.151	0.116	0.066	0.285	0.158
1.5	0.502	0.571	0.300	0.112	0.031	0.042	0.031	0.019	0.075	0.042
1.75	0.247	0.344	0.150	0.052	0.014	0.013	0.009	0.006	0.019	0.013
2	0.119	0.215	0.076	0.029	0.007	0.005	0.003	0.002	0.006	0.005

Table 1A:. Reference genomes are hg19 (H.s.), mm10 (M.m.), dm3 (D.m.), ce10 (C.e.), ddAX4 (D.d.), sacCer3 (S.c.), NCB.I20020305 (S.p.), NCBI.20070724 (P.f.), Ecol.NCBI.20080805 (E.c.) and TAIR9 (A.t.). Note that the E.c. reference genome contains 13 genomes of different *E. coli* strains.

Threshold	Macaque	Chimp	Cow	Pig	Dog	Rat	Chicken	Zebrafish	Bee	Rice
1	2.425	2.391	2.508	2.524	2.791	2.457	2.079	1.137	1.281	2.279
1.25	0.995	0.989	1.132	1.181	1.340	1.083	0.836	0.412	0.639	1.029
1.5	0.478	0.486	0.615	0.651	0.739	0.576	0.395	0.190	0.370	0.516
1.75	0.224	0.238	0.327	0.363	0.412	0.325	0.196	0.091	0.213	0.254
2	0.101	0.114	0.174	0.199	0.228	0.189	0.101	0.047	0.116	0.119

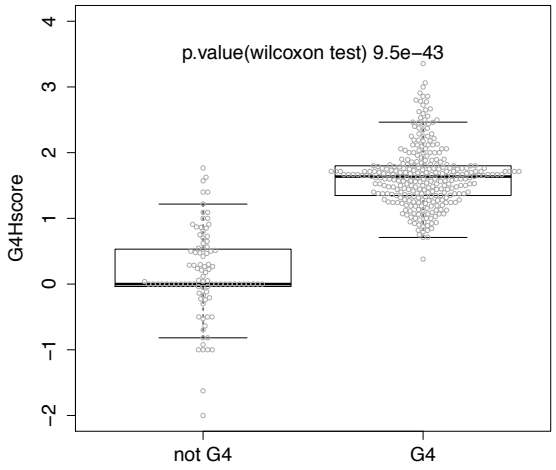
Table 1B: Reference genomes are rheMac3 (Mac), panTro3 (Chimp), bosTau6 (Cow), susScr3 (Pig), canFam3 (Dog), rn5 (Rat), galGal4 (Chicken), danRer7 (Zebrafish), apiMel2 (Bee) and MSU7 (rice).

Table 2: G4FS in promoter region of the human genome

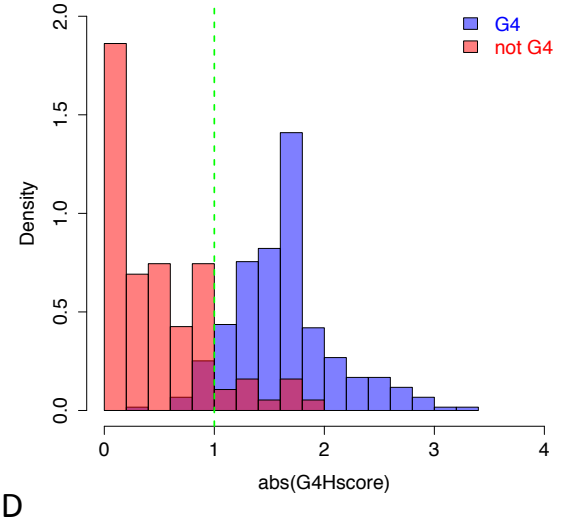
	G4H1	G4H1.2	G4H1.5	G4H1.75	G4H2	QP37	QP312	QP27
G4FS list size (no chrM)	*6938933*	*3674822*	*1436253*	*707090*	*339975*	*361982*	*706788*	*8572750*
Promoter with G4FS≥1 (%)	**94.1*****	**84.5*****	**66.5*****	**51.9*****	**36.9*****	**38.7*****	**52.7*****	**94.2**
***Enrichment** in promoter with G4FS≥1*	*1.0*	*1.1*	*1.5*	*2.0*	*2.6*	*2.6*	*2.0*	*1.0*
(resampled background, mean±sd%)	*(92.8±0.1)*	*(76.8±0.2)*	*(45.4±0.3)*	*(26.6±0.2)*	*(14.0±0.2)*	*(14.9±0.2)*	*(26.8±0.2)*	*(95.6±0.1)*
Fraction GFS in Promoter (%)	**2.8*****	**3.5*****	**4.7*****	**5.6*****	**6.5*****	**6.8*****	**6.2*****	**2.8*****
Enrichment vs. Global	*2.2*	*2.8*	*3.8*	*4.5*	*5.2*	*5.5*	*5.0*	*2.3*
(global background,%)	*(1.2)*	*(1.2)*	*(1.2)*	*(1.2)*	*(1.2)*	*(1.2)*	*(1.2)*	*(1.2)*
Enrichment vs. Resampled	*2.0*	*2.5*	*3.1*	*3.7*	*4.3*	*4.5*	*4.1*	*2.0*
(resampled background, mean±sd%)	*(1.4±0.004)*	*(1.4±0.006)*	*(1.5±0.01)*	*(1.5±0.014)*	*(1.5±0.021)*	*(1.5±0/021)*	*(1.5±0.014)*	*(1.4±0.004)*
G4FS/kb in Promoter	**5.38*****	**3.62*****	**1.88*****	**1.12*****	**0.62*****	**0.69*****	**1.23*****	**6.78*****
Enrichment vs. Global	*2.2*	*2.8*	*3.8*	*4.5*	*5.2*	*5.5*	*5.0*	*2.3*
(global background)	*(2.43)*	*(1.28)*	*(0.50)*	*(0.25)*	*(0.12)*	*(0.13)*	*(0.25)*	*(3.00)*
Enrichment vs. Resampled	*2.0*	*2.5*	*3.2*	*3.8*	*4.4*	*4.6*	*4.1*	*2.0*
(resampled background, mean±sd)	*(2.71±0.008)*	*(1.46±0.006)*	*(0.59±0.004)*	*(0.29±0.003)*	*(0.14±0.002)*	*(0.15±0.002)*	*(0.3±0.003)*	*(3.39±0.01)*

The size of the different G4FS list (G4Hunter with threshold of 1, 1.2, 1.5, 1.75 and 2, and Quarparser QP37, QP312 and QP312) was computed excluding the mitochondrial DNA. Fraction of promoter with at least one GFS was computed first by combining all the promoter regions of the Known Genes from UCSC (TSS-1000 to TSS) and then looking the presence of at least one G4FS in these regions. The enrichment was calculated using a resampled background. Fraction of G4FS in promoter was computed by counting the occurrence of G4FS in a promoter region as defined before divided by the size of the G4FS list. G4FS density (G4FS/kb) was computed by dividing the number of G4FS found in promoter with sum of the length of the promoter regions in kb. For these two metrics, enrichment was computed compared to a global background or to a resampled background (see main text for details).
*** means a p-value < 1/1000 for the null hypothesis: G4FS are randomely distributed on the chromosoms.

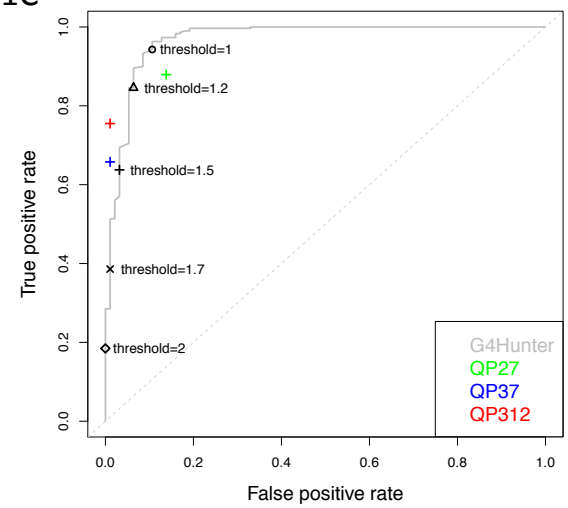
1A


p.value(wilcoxon test) 9.5e−43
G4Hscore
not G4
G4

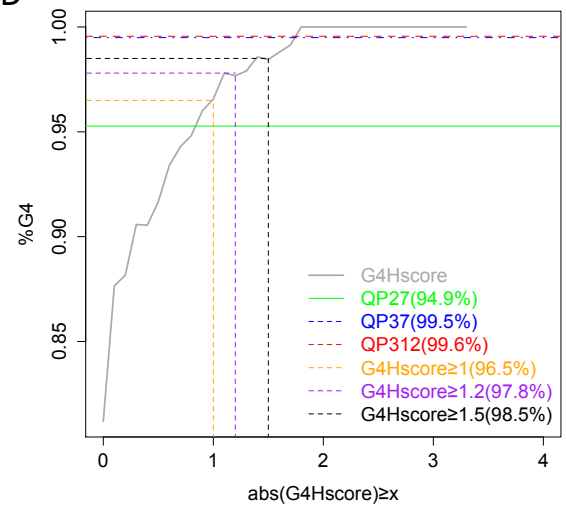
1B


G4
not G4
Density
abs(G4Hscore)

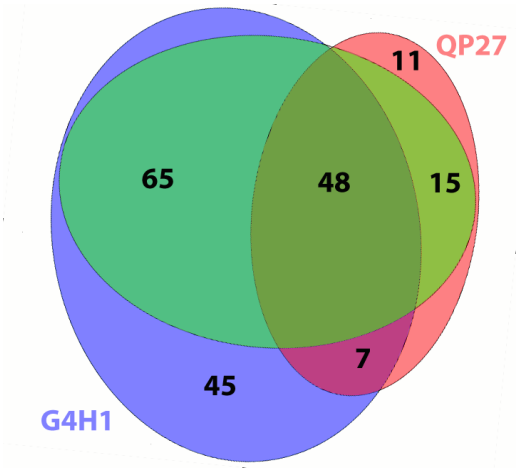
1C


threshold=1
threshold=1.2
threshold=1.5
threshold=1.7
threshold=2
G4Hunter
QP27
QP37
QP312
True positive rate
False positive rate

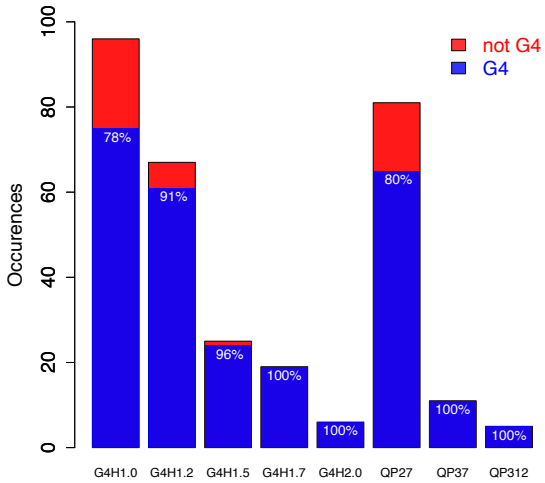
1D


G4Hscore
QP27(94.9%)
QP37(99.5%)
QP312(99.6%)
G4Hscore≥1(96.5%)
G4Hscore≥1.2(97.8%)
G4Hscore≥1.5(98.5%)
%G4
abs(G4Hscore)≥x

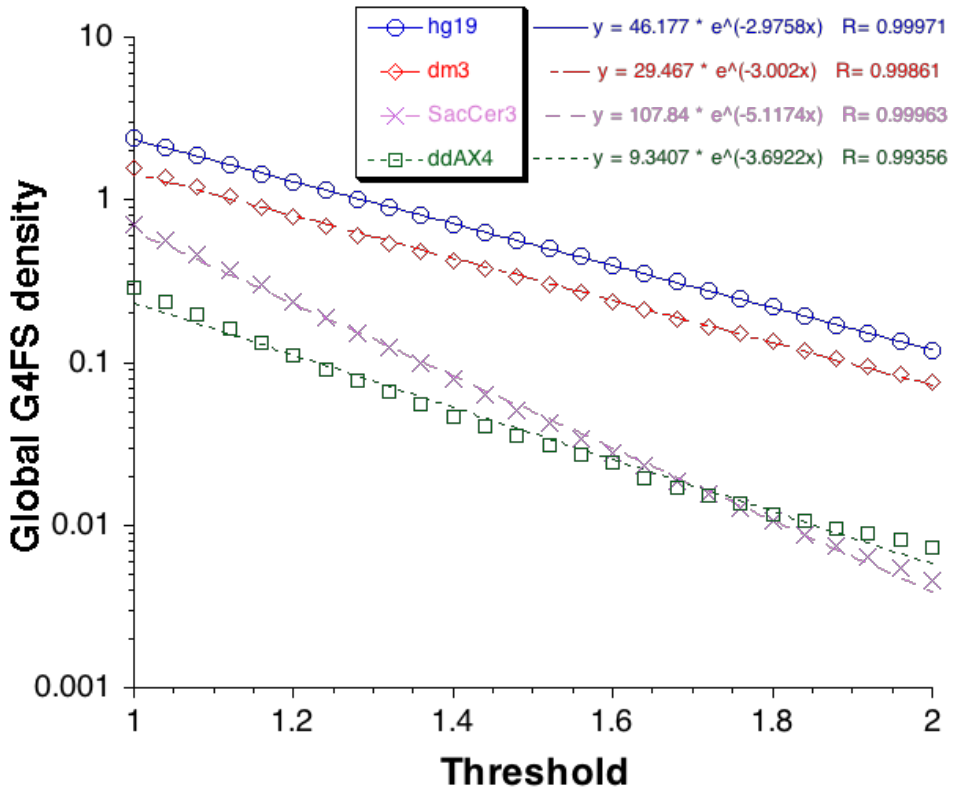
2A


QP27
11
65
48
15
7
45
G4H1

2B


not G4
G4
Occurences
0
20
40
60
80
100
78%
91%
96%
100%
100%
80%
100%
100%
G4H1.0
G4H1.2
G4H1.5
G4H1.7
G4H2.0
QP27
QP37
QP312

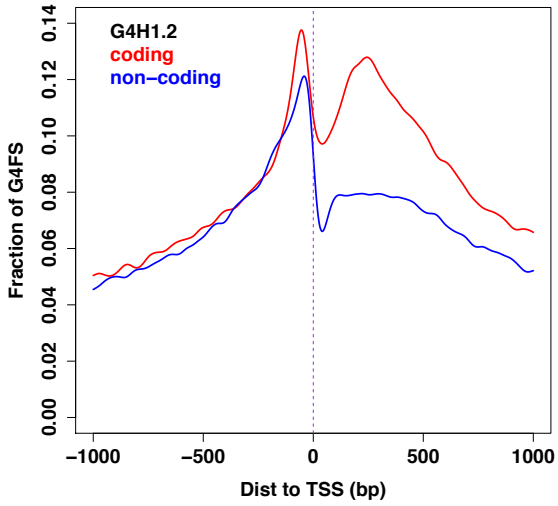
3

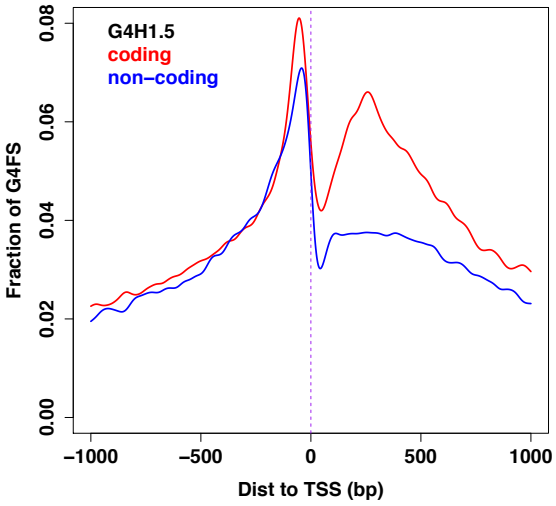

hg19
dm3
SacCer3
ddAX4
y = 46.177 * e^(-2.9758x) R= 0.99971
y = 29.467 * e^(-3.002x) R= 0.99861
y = 107.84 * e^(-5.1174x) R= 0.99963
y = 9.3407 * e^(-3.6922x) R= 0.99356
Global G4FS density
10
1
0.1
0.01
0.001
1
1.2
1.4
1.6
1.8
2
Threshold

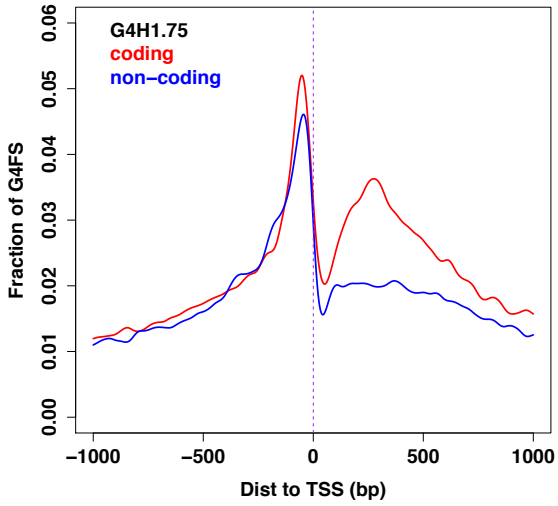
4

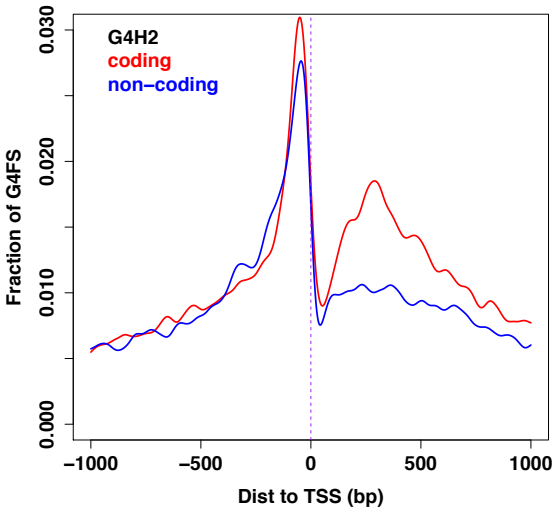

Human hg19
chr8
chr8:128,746,310-128,750,317
Go
4 000 bp
RefSeq Genes
MYC
G4H1.2.bed
G4H1.5.bed
G4H1.75.bed
G4H2.bed
QP27.bed
QP312.bed
QP37.bed
Human hg19
chr8
chr8:128,650,294-128,850,724
Go
200 kb
RefSeq Genes
MYC
MIR1204
G4FS/kb (G4H1.20)
G4FS/kb (G4H1.5)
G4FS/kb (G4H1.75)
G4FS/kb (G4H2)
G4FS/kb (QP27
G4FS/kb (QP312)
G4FS/kb (QP37)

5

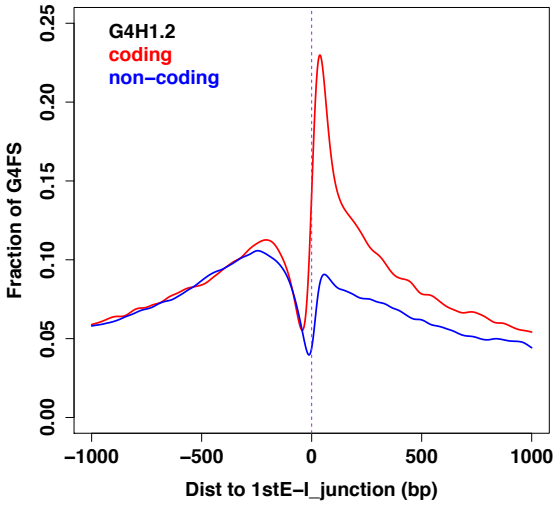

G4H1.2
coding
non−coding
Fraction of G4FS
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
−1000
−500
0
500
1000
Dist to TSS (bp)

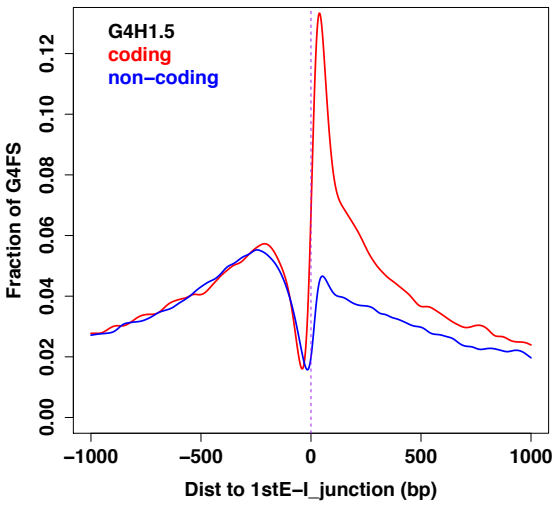

G4H1.5
coding
non−coding
Fraction of G4FS
0.00
0.02
0.04
0.06
0.08
−1000
−500
0
500
1000
Dist to TSS (bp)

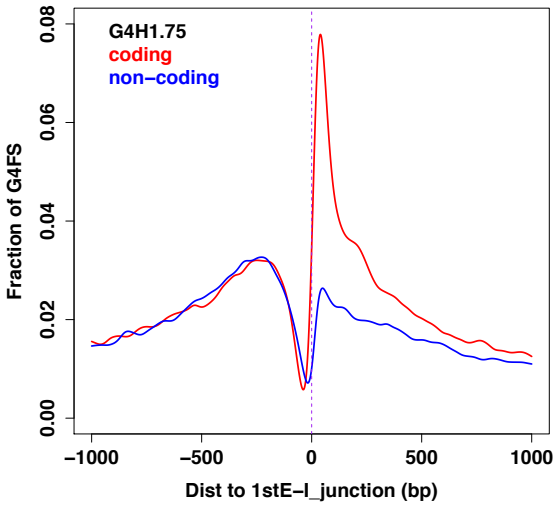

G4H1.75
coding
non−coding
Fraction of G4FS
0.00
0.01
0.02
0.03
0.04
0.05
0.06
−1000
−500
0
500
1000
Dist to TSS (bp)

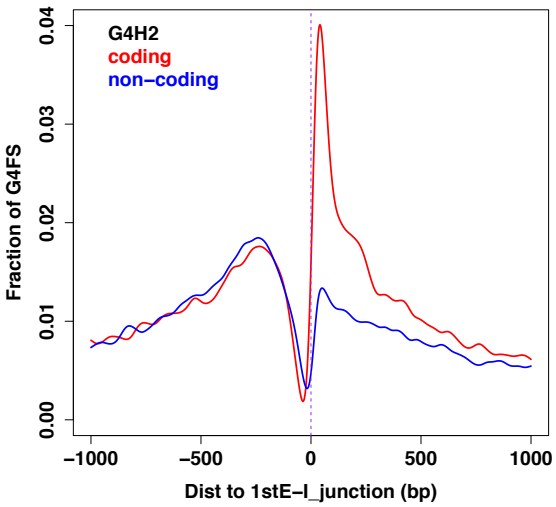

G4H2
coding
non−coding
Fraction of G4FS
0.000
0.010
0.020
0.030
−1000
−500
0
500
1000
Dist to TSS (bp)

6

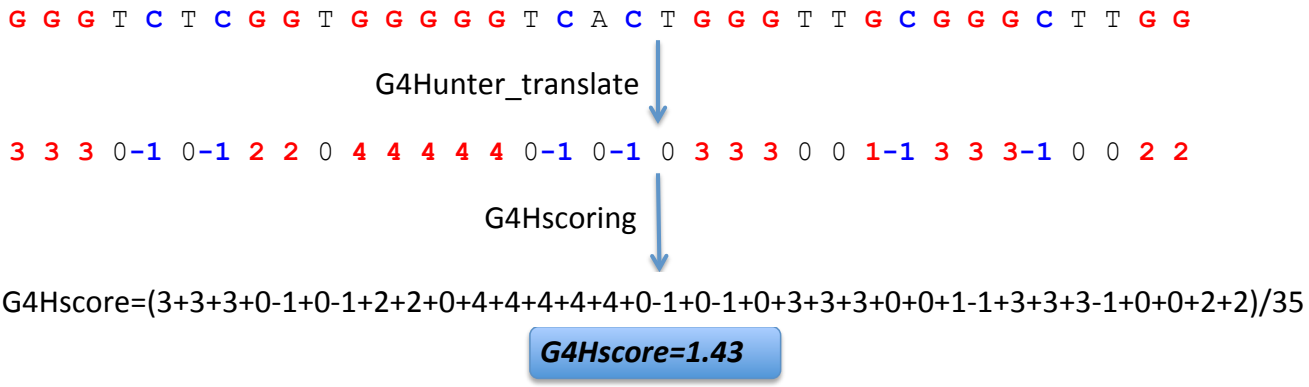

G4H1.2
coding
non−coding
Fraction of G4FS
0.00
0.05
0.10
0.15
0.20
0.25
−1000
−500
0
500
1000
Dist to 1stE−I_junction (bp)


G4H1.5
coding
non−coding
Fraction of G4FS
0.00
0.02
0.04
0.06
0.08
0.10
0.12
−1000
−500
0
500
1000
Dist to 1stE−I_junction (bp)


G4H1.75
coding
non−coding
Fraction of G4FS
0.00
0.02
0.04
0.06
0.08
−1000
−500
0
500
1000
Dist to 1stE−I_junction (bp)


G4H2
coding
non−coding
Fraction of G4FS
0.00
0.01
0.02
0.03
0.04
−1000
−500
0
500
1000
Dist to 1stE−I_junction (bp)

S1A



Classical examples:

21g	GGGTTAGGGTTAGGGTTAGGG	1.71	
A22g	AGGGTTAGGGTTAGGGTTAGGG	1.64	
cMYC	TGAGGGTGGGTAGGGTGGGTAA	1.68	
27Kras	GGGCGGTGTGGGAAGAGGGAAGAGGGG	1.81	
cKit1	GGGAGGGCGCTGGGAGGAGGG	1.86	
cKit2GG	GGGCGGGCGCGAGGGAGGGG	2.1	
Oxy28	GGGGTTTTGGGGTTTTGGGGTTTTGGGG	2.29	
22Agm4	ATGGTTAGTGTTAGGTTTAGTG		0.55
CGG12	(CGG)12	1	
ds26	CAATCGGATCGAATTCGATCCGATTG	0	
21Ctel	CCCTAACCCTAACCCTAACCC	-1.71	

S1B

G4Hunter on a long sequence

GAGGACAAGGAGGTGCGAGGAAAGGGGTTGGGGGATGGTCCCACAGGCAGCCACACCTGAGGCGTGGGCGGCCGGTAGGAGCTGGGGGAGGGCG
GGGAGAAGAGGGGTTTCTGTGTGTAGA

(1) G4Hunter_translate

1 0 2 2 0 -1 0 0 2 2 0 2 2 0 1 -1 1 0 2 2 0 0 0 4 4 4 4 0 0 4 4 4 4 4 4 0 0 2 2 0 -3 -3 -3 0 -1 0 2 2 -1 0 1 -2 -2 0 -1 0 -2 -2 0
1 0 2 2 -1 1 0 3 3 3 -1 2 2 -2 -2 2 2 0 0 2 2 0 1 -1 0 4 4 4 4 4 4 0 3 3 3 -1 4 4 4 4 0 1 0 0 1 0 4 4 4 4 0 0 0 -1 0 1 0 1 0 1
0 0 1 0

(2) G4Hscore_windowed (k=25)

1.00 1.12 1.28 1.20 1.12 1.28 1.48 **1.64 1.80 1.88 1.80 1.80 1.80 1.80 1.80 1.64 1.56** 1.40 1.40 1.28 1.20 1.28 1.36 1.32 1.16
1.04 0.80 0.56 0.56 0.52 0.36 0.12 -0.12 -0.28 -0.40 -0.40 -0.32 -0.32 -0.44 -0.40 -0.28 -0.04 0.20 0.32 0.32 0.40 0.40 0.24 0.20 0.28
0.32 0.40 0.48 0.56 0.68 0.68 0.80 0.84 0.84 0.96 1.12 1.20 1.28 1.48 1.44 **1.56 1.56 1.56** 1.40 **1.60 1.68 1.76 2.00 2.08 2.04**
1.96 1.96 2.00 1.92 2.00 2.16 2.28 2.48 2.48 2.32 2.16 1.96 1.80 1.68 1.68 1.60 1.48 1.40 1.44 1.28 1.16 1.00

(3) Window extraction (threshold=1.5)

	start	end	width	G4Hscore (window=25)
[1]	8	17	10	[1.64 1.80 1.88 1.80 1.80 1.80 1.80 1.80 1.64 1.56]
[2]	66	68	3	[1.56 1.56 1.56]
[3]	70	91	22	[1.60 1.68 1.76 2.00 2.08 2.04 1.96 1.96 2.00 1.92 2.00 2.16 2.28 2.48 2.48 2.32 2.16 1.96 1.80 1.68 1.68 1.60]

(4) Sequence extraction

	start	end	width	sequence	G4Hscore
[1]	8	41	34	[AGGAGGTGCGAGGAAAGGGGTTGGGGGATGGTCC]	1.44
[2]	66	92	27	[GGGC**GGCCGGTAGGAGCTGGGGGAGGG**]	1.67
[3]	70	115	46	[**GGCCGGTAGGAGCTGGGGGAGGG**CGGGGAGAAGAGGGGTTTCTGTG]	1.54

(5) Fusion of overlapping sequences

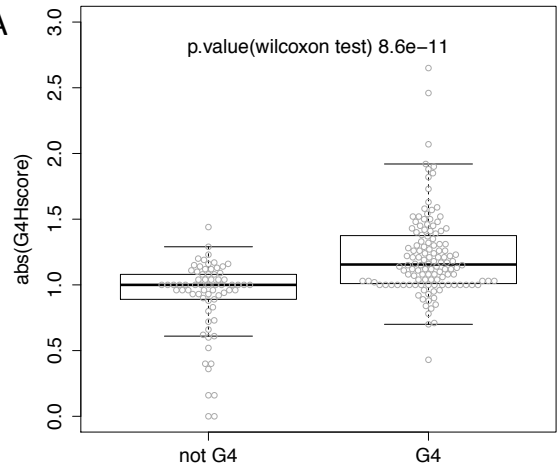
	start	end	width	sequence	G4Hscore
[1]	8	41	34	[**A**GGAGGTGCGAGGAAAGGGGTTGGGGGATGG**TCC**]	1.44
[2]	66	115	50	[GGGCGGCCGGTAGGAGCTGGGGGAGGGCGGGGAGAAGAGGGGTTTCTGTG]	1.58

(6) Extremity repair and final G4Hscore computation

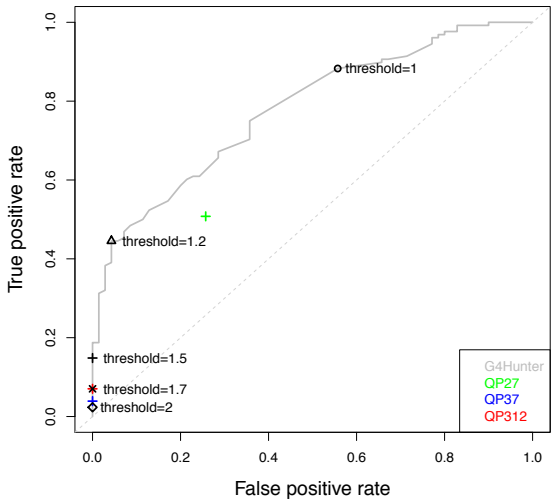
	start	end	sequence	G4Hscore
[1]	9	38	GGAGGTGCGAGGAAAGGGGTTGGGGGATGG	1.77
[2]	66	115	GGGCGGCCGGTAGGAGCTGGGGGAGGGCGGGGAGAAGAGGGGTTTCTGTG	1.58

GAGGACA**A**GGAGGTGCGAGGAAAGGGGTTGGGGGATGG**TCC**CACAGGCAGCCACACCTGAGGCGT**GGGCGGCCGGTAGGAGCTGGGGGAGGGCGGGGAGAAGAGGGGTTT
CTGTGTG**TAGA

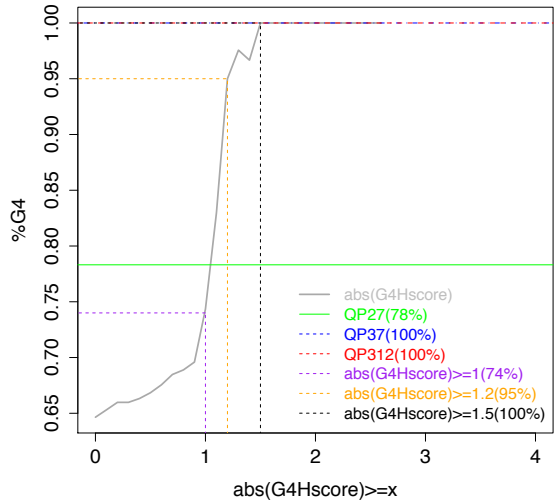
S2A



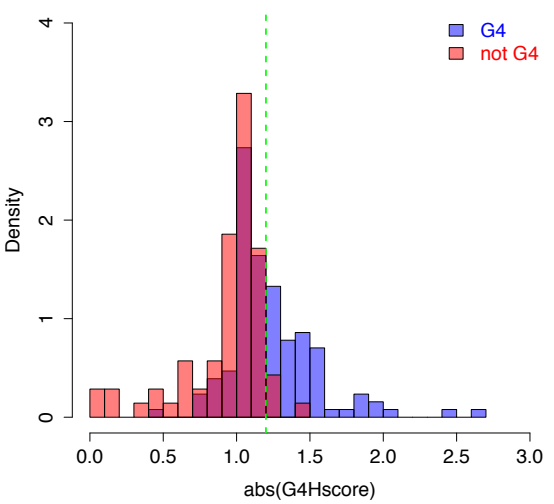
S2B



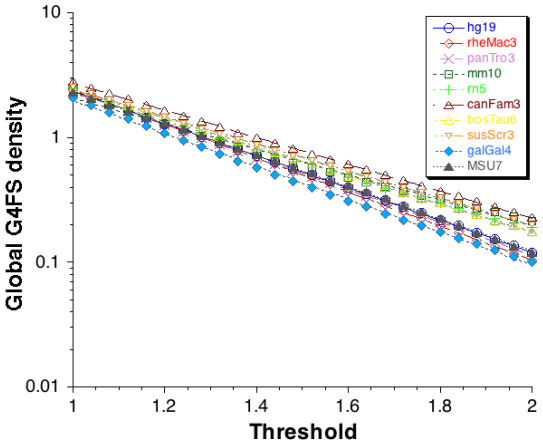
S2C



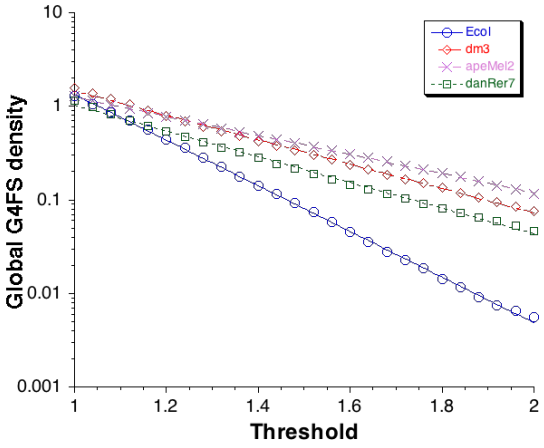
S2D



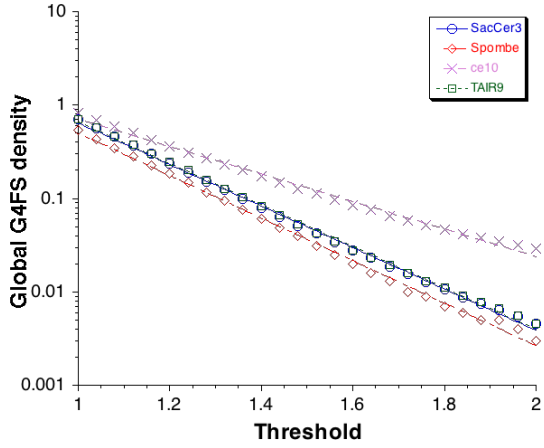
S3A



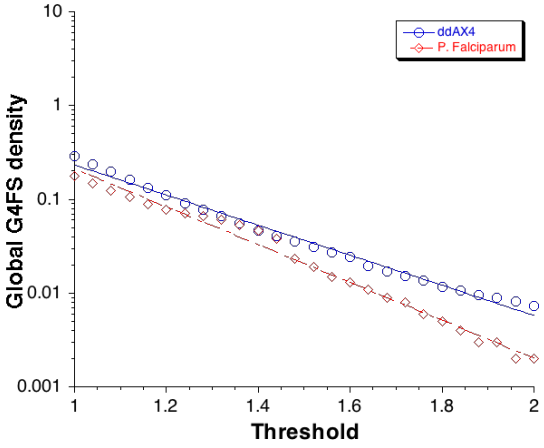
S3B



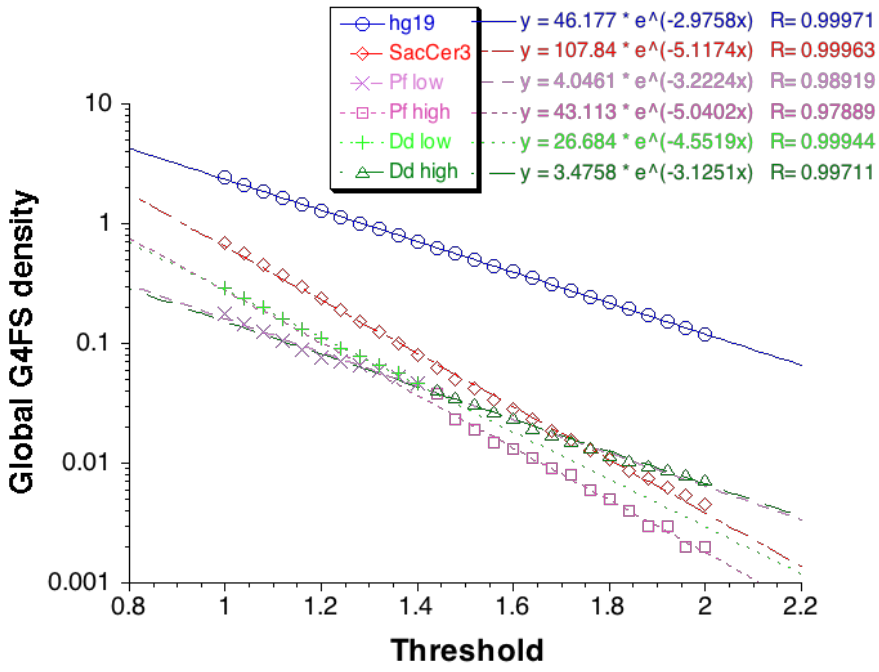
S3C



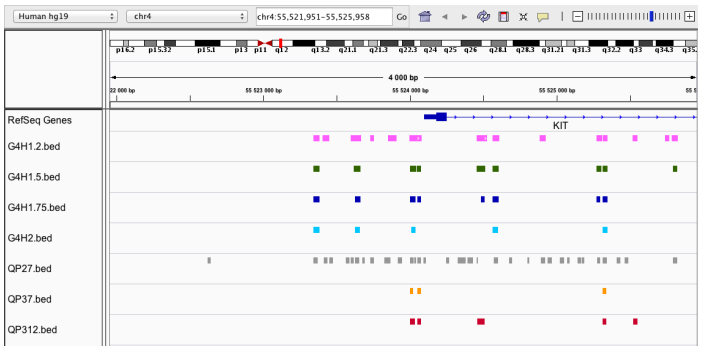
S3D



S3E


hg19
SacCer3
Pf low
Pf high
Dd low
Dd high
y = 46.177 * e^(-2.9758x) R= 0.99971
y = 107.84 * e^(-5.1174x) R= 0.99963
y = 4.0461 * e^(-3.2224x) R= 0.98919
y = 43.113 * e^(-5.0402x) R= 0.97889
y = 26.684 * e^(-4.5519x) R= 0.99944
y = 3.4758 * e^(-3.1251x) R= 0.99711
Global G4FS density
10
1
0.1
0.01
0.001
0.8
1
1.2
1.4
1.6
1.8
2
2.2
Threshold

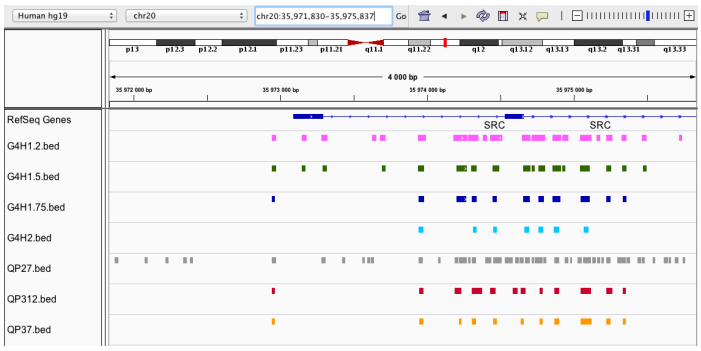
S4A


Human hg19
chr4
chr4:55,521,951–55,525,958
4 000 bp
RefSeq Genes
KIT
G4H1.2.bed
G4H1.5.bed
G4H1.75.bed
G4H2.bed
QP27.bed
QP37.bed
QP312.bed


Human hg19
chr18
chr18:60,984,608–60,988,615
4 000 bp
RefSeq Genes
BCL2
BCL2
G4H1.2.bed
G4H1.5.bed
G4H1.75.bed
G4H2.bed
QP27.bed
QP312.bed
QP37.bed

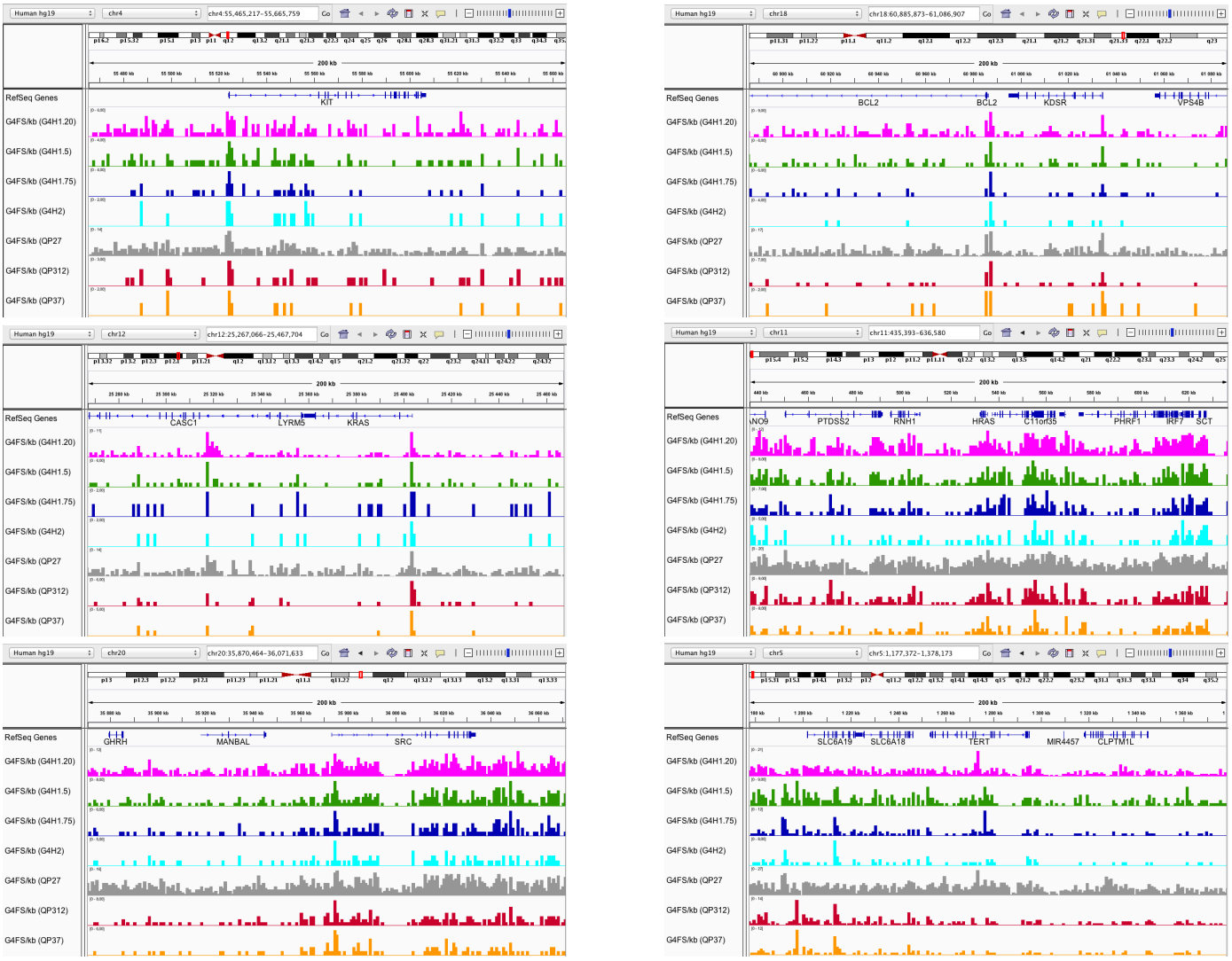

Human hg19
chr12
chr12:25,401,850–25,405,857
4 000 bp
RefSeq Genes
KRAS
G4H1.2.bed
G4H1.5.bed
G4H1.75.bed
G4H2.bed
QP27.bed
QP312.bed
QP37.bed


Human hg19
chr11
chr11:533,545–537,552
4 000 bp
RefSeq Genes
HRAS
LRR
G4H1.2.bed
G4H1.5.bed
G4H1.75.bed
G4H2.bed
QP27.bed
QP312.bed
QP37.bed

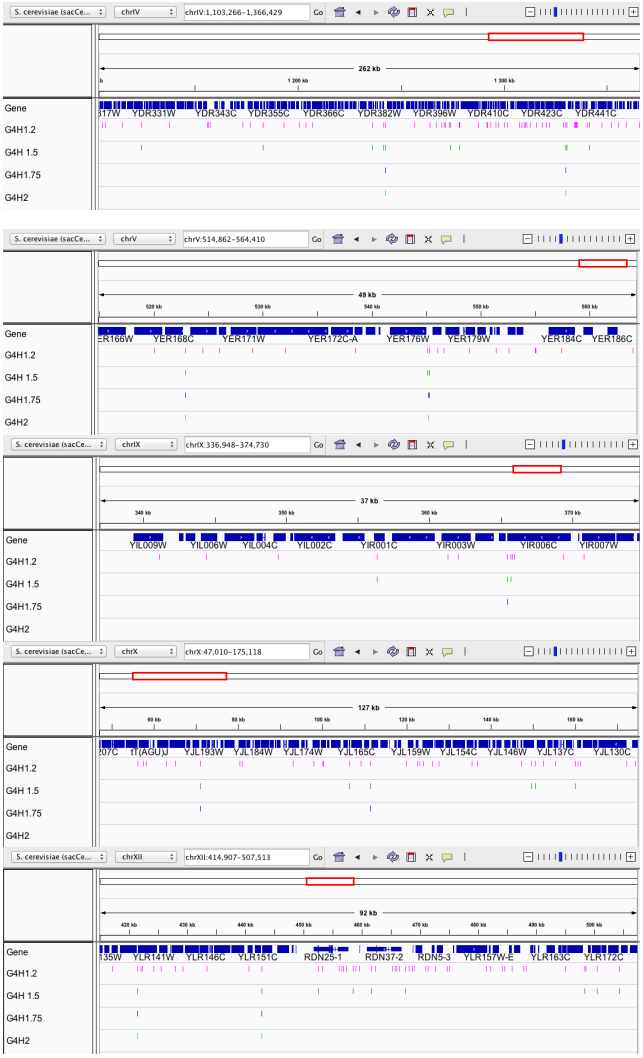

Human hg19
chr20
chr20:35,971,830–35,975,837
4 000 bp
RefSeq Genes
SRC
SRC
G4H1.2.bed
G4H1.5.bed
G4H1.75.bed
G4H2.bed
QP27.bed
QP312.bed
QP37.bed


Human hg19
chr5
chr5:1,293,158–1,297,165
4 000 bp
RefSeq Genes
TERT
G4H1.2.bed
G4H1.5.bed
G4H1.75.bed
G4H2.bed
QP27.bed
QP312.bed
QP37.bed

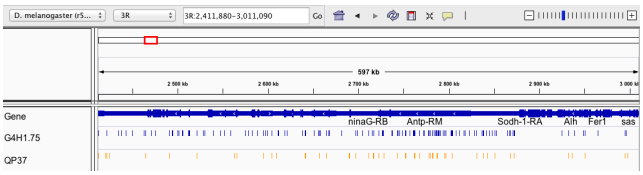
S4B

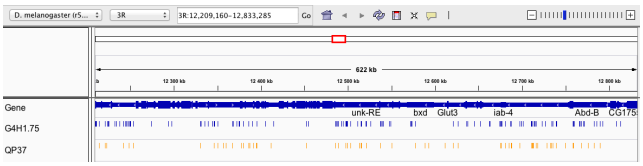

Human hg19
chr4
chr4:55,465,217-55,665,759
RefSeq Genes
KIT
G4FS/kb (G4H1.20)
G4FS/kb (G4H1.5)
G4FS/kb (G4H1.75)
G4FS/kb (G4H2)
G4FS/kb (QP27)
G4FS/kb (QP312)
G4FS/kb (QP37)
chr18
chr18:60,885,873-61,086,907
BCL2
BCL2
KDSR
VPS4B
chr12
chr12:25,267,066-25,467,704
CASC1
LYRM5
KRAS
chr11
chr11:435,393-636,580
NO9
PTDSS2
RNH1
HRAS
C11orf35
PHRF1
IRF7
SCT
chr20
chr20:35,870,464-36,071,633
GHRH
MANBAL
SRC
chr5
chr5:1,177,372-1,378,173
SLC6A19
SLC6A18
TERT
MIR4457
CLPTM1L

S4C



SacCer3

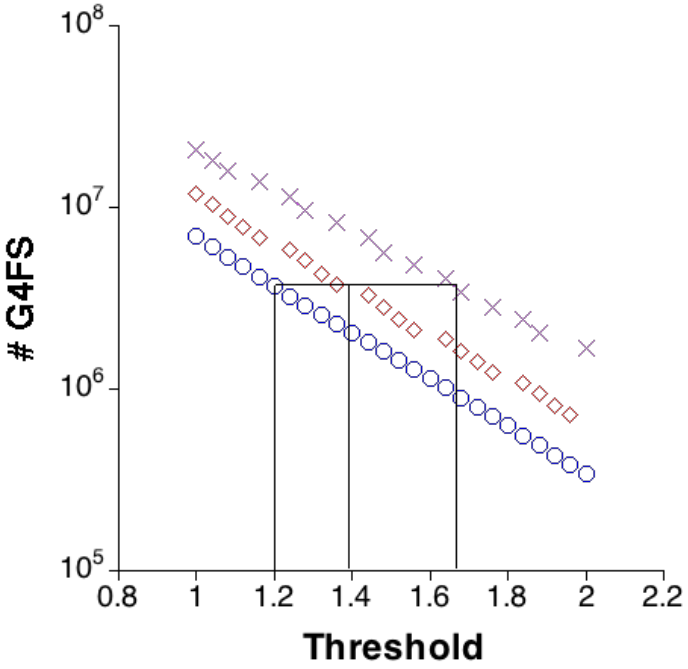




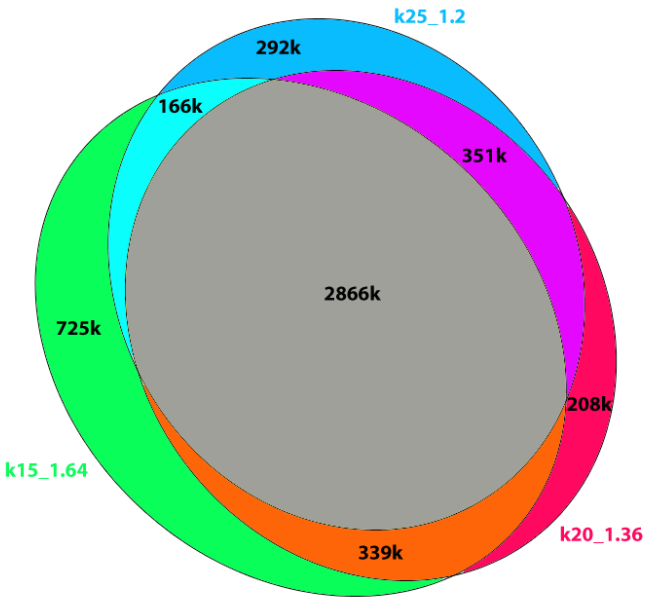
dm3

rDNA(hg19)

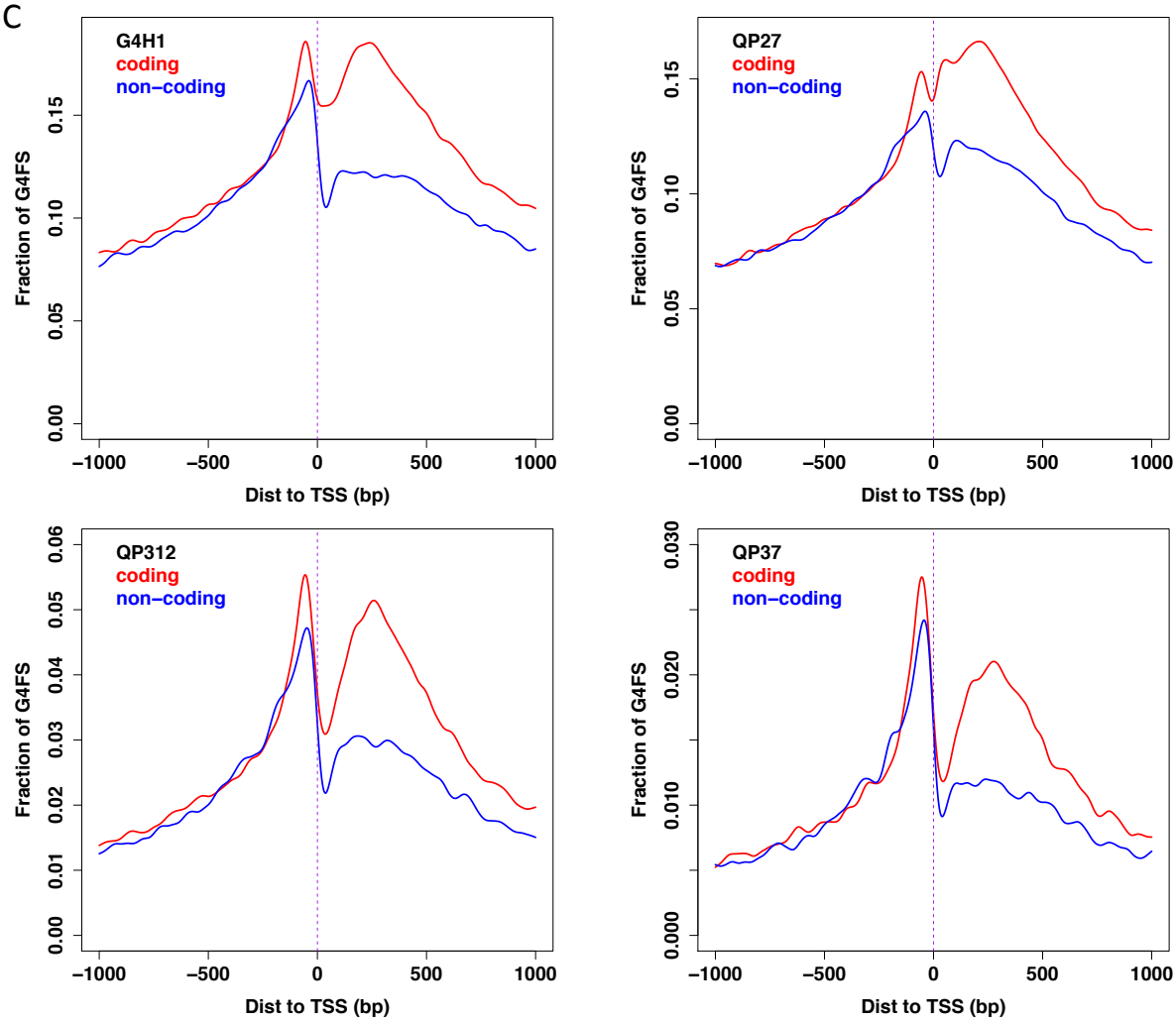
S5A


10^8
10^7
10^6
10^5
G4FS
0.8
1
1.2
1.4
1.6
1.8
2
2.2
Threshold

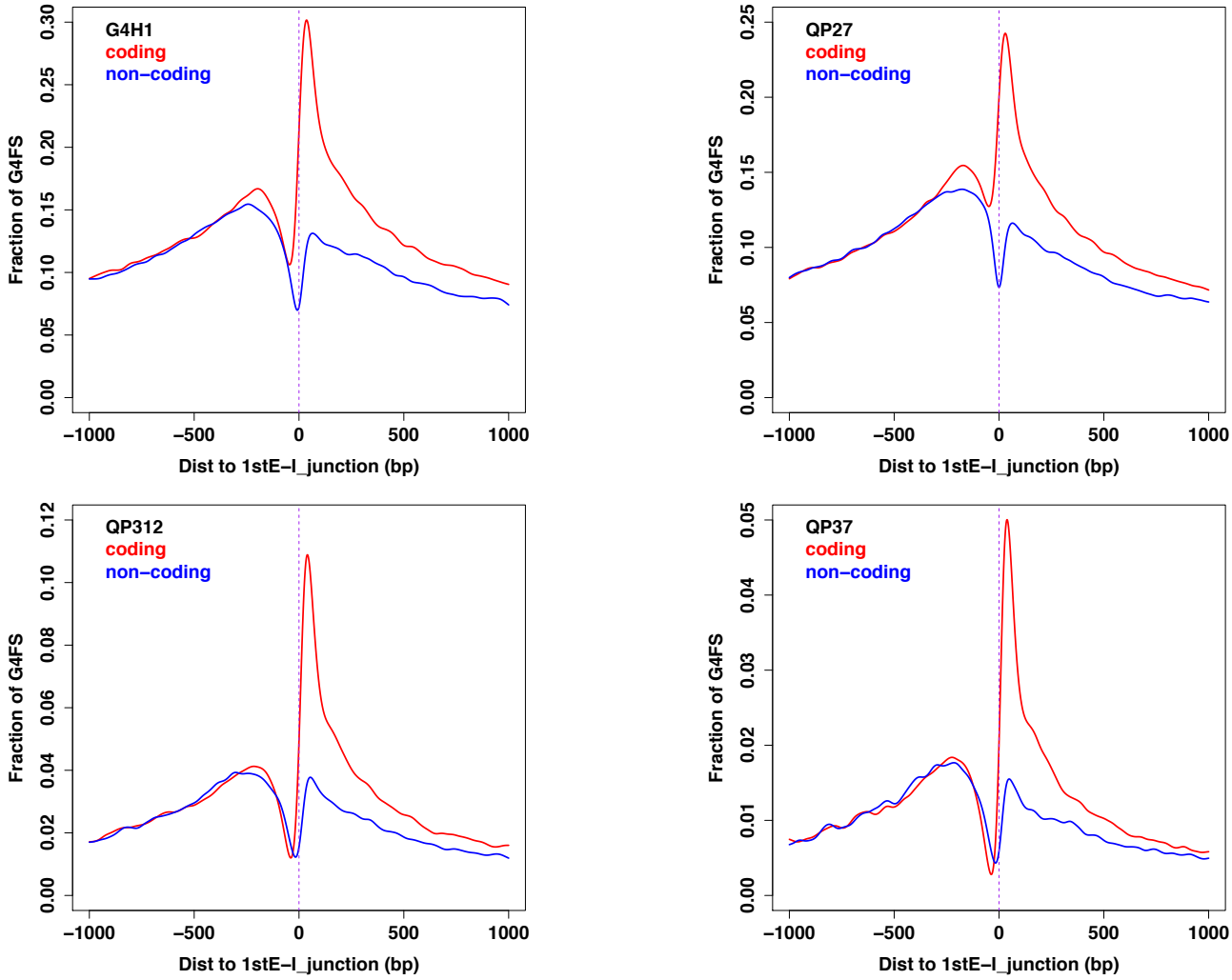
S5B


k25_1.2
292k
166k
351k
2866k
725k
208k
k15_1.64
339k
k20_1.36

S5C


G4H1
coding
non−coding
Fraction of G4FS
0.00
0.05
0.10
0.15
−1000
−500
0
500
1000
Dist to TSS (bp)
QP27
coding
non−coding
Fraction of G4FS
0.00
0.05
0.10
0.15
−1000
−500
0
500
1000
Dist to TSS (bp)
QP312
coding
non−coding
Fraction of G4FS
0.00
0.01
0.02
0.03
0.04
0.05
0.06
−1000
−500
0
500
1000
Dist to TSS (bp)
QP37
coding
non−coding
Fraction of G4FS
0.000
0.010
0.020
0.030
−1000
−500
0
500
1000
Dist to TSS (bp)

S6


G4H1
coding
non−coding
Fraction of G4FS
0.00
0.05
0.10
0.15
0.20
0.25
0.30
−1000
−500
0
500
1000
Dist to 1stE−I_junction (bp)
QP27
coding
non−coding
Fraction of G4FS
0.00
0.05
0.10
0.15
0.20
0.25
−1000
−500
0
500
1000
Dist to 1stE−I_junction (bp)
QP312
coding
non−coding
Fraction of G4FS
0.00
0.02
0.04
0.06
0.08
0.10
0.12
−1000
−500
0
500
1000
Dist to 1stE−I_junction (bp)
QP37
coding
non−coding
Fraction of G4FS
0.00
0.01
0.02
0.03
0.04
0.05
−1000
−500
0
500
1000
Dist to 1stE−I_junction (bp)

Supplementary Information for
Reevaluation of quadruplex propensity with G4Hunter

Amina Bedrat[1,2], Laurent Lacroix[3]* & Jean-Louis Mergny[1,2]*

1. Université de Bordeaux, ARNA Laboratory, F-33000 Bordeaux, France;
2. Inserm U869, IECB, F-33600 Pessac, France;
3. CNRS-Université de Toulouse UMR5099, F-31000, Toulouse, France.

Discussion of the mitochondrial results:

Based on the conclusion from the biophysical tests, the sequences were classified as G4 (n=128) and not G4 (n=70) and subject to the same statistical analysis as the reference dataset (see supplementary Figure S2). The average G4Hscore was 1.22±0.31 for the G4 class and 0.91±0.29 for the not G4 class (supplementary Figure S2A). The difference is less pronounced than for the reference dataset but still significatif according to a non-parametric Wilcoxon rank-sum/ Mann-Whitney U-test (null hypothesis: distributions are not different), with a p-value of $8.6 \cdot 10^{-11}$. The discriminating score appears to be 1.2 (supplementary Figure S2D) but G4Hscore are less dispersed as the one from the reference dataset as most of the sequences have been selected with a threshold of 1.

As for the reference dataset, a ROC analysis (supplementary Figure S2B) illustrates that G4Hunter perform better than a random estimator on this dataset, but the accuracy reflected by the area under the curve AUC=0.78) is less than for the reference dataset, which by construction was biased for G4FS. From this analysis, G4Hunter appears to perform better the Quadparser fur G-runs of 2 or more as the ROC curve is above this quadparser point (green cross). The ROC analysis allow us to propose that a threshold of 1.2 (triangle on supplementary Figure S2B) appears as an excellent compromise between true and false positive rate as it is the point the farest from the diagonal. This threshold allow a similar TPR than Quadparser for runs of 2 Gs (0.45 vs. 0.51) but with a much reduced FPR (0.04 vs. 0.26). Using Quadparser with runs of Gs of 3 or more allows a FPR of 0, comparable to G4Hunter with a threshold of 1.5 or more but with a dramatically reduced TPR (less than 0.04 for Quadparser vs. 0.15 for G4Hunter with a threshold of 1.5).

The ROC curve analysis allows to evaluate the ability of the G4Hscore to determine which sequence form indeed a G4 in the whole population of G4 in the mitochondria genome, but this does not allow to determine the precision of the G4Hscore. This correspond to the fraction of hits called by G4Hunter or Quadparser that forms G4. This precision is of 78%, 100% and 100% for QP27, QP37 and QP312 respectively. For sequences which G4Hscore is above 1, the precision is 74% but it rise up to 95% with a threshold of 1.2. With a threshold above 1.5, 100% of the hits form quadruplexes (supplementary Figure S2C).

Supplementary Figure Legends

Figure S1:
Principle of the scoring procedure
A- Decomposition of the scoring procedure for a single sequence. The nucleotide sequence is first translate number between -4 and +4 using the G4Hunter algorithm and then the G4Hscore is calculated by taking the average of these numbers. Below are represented some classical examples using the following color code (G4 in red, i-motif in blue, not G4 (nor i-motif) in black).
B- G4Hunter procedure to extract G4FS from long sequences. *(1)* The nucleotide sequence is first translate in number between -4 and +4 unsing the G4Hunter algorithm (G4Hunter_translate). *(2)* The G4Hscore is then computed on moving window of size k (runmean in **R**, k=25). *(3)* Regions were this runmean of above the threshold (here 1.5) are selected and *(4)* the corresponding sequences retrieved. The G4Hscore of these sequence is calculated as in supplementary Figue S1A. *(5)* Overlapping sequences are fused to avoid multiple counting ot the same G4FS (overlapping sequences are indicated in blue) and the new G4Hscore is calculated. *(6)* extremity of the sequences are process to eliminate non G (or C) bases (in green) and also to retrieve from the initial sequence G (or C) that would have been cut-out by the windowing procedure (not shown in this example). Note that between step *(3)* and *(4)*, the extracted window sizes have been corrected using the runmean window size and thus the last two aeras that were not overlapping at the runmean level are now overlapping at the nucleotide level.

Figure S2:
A- boxplot of the absolute value of the G4Hscore for the mitochondria dataset. Open circle represents the individual G4Hscore values for the two classes: G4 and not G4. P-value for a Wilcoxon test between the 2 classes is indicated.
B- ROC curve for G4Hunter on the reference dataset. Black symbols represent the position of individual threshold values for G4Hunter. Green, blue and red crosses represent position of the corresponding ROC values after applying QuadParser algorithm on the reference dataset with the following settings: runs of 2Gs and loops length between 1 and 7 (QP27, green), runs of 3Gs and loops length between 1 and 7 (QP37, blue) and runs of 3Gs and loops length between 1 and 12 (QP312, red).
C- Precision *vs*. threshold for G4Hunter on the mitochondria dataset. Fraction of sequences classified as G4 forming and which abs(G4Hscore) is above the threshold on the X-axis. Precision for the threshold 1, 1.2 and 1.5 are indicated with dotted line in purple, orange and black respectively. Precision with QP27, QP37 and QP312 are indicated in green, blue and red respectively.
D- Histogram of density distribution of the absolute values of the G4Hscore for the two classes of the mitochondria dataset (Blue: G4 forming class, Red: not G4 forming class). The green dotted line indicates the value of abs(G4Hscore) for which more G4 than not G4 are found in this density histogram.

Figure S3:
Global density of hits per kb found by G4Hunter at different threshold between 1 and 2 and exponential fits for the whole genomes.
A- G4FS rich genomes: hg19 (blue), rheMac3 (red), panTro3 (violet), mm10 (dark green), rn5 (green), canFam3 (brown), bosTau6 (yellow), susScr3 (orange), galGal4 (light blue), MSU7 (grey);
B- intermediate G4FS genome: *E. coli* (blue), fly (red), bee (violet) and zebrafish (green);

C- low G4FS genomes: budding (blue) and fission (red) yeasts, *C. elegans* (violet) and *A. thaliana* (green);
D- very poor G4FS genomes: *P. falciparum* (red) and *D. discoideum* (blue).
E- Proposed decomposition for the fit of the very poor G4FS genomes (pink: *P. falciparum* and green: *D. discoideum*) for high and low thresholds compared to the human (blue) and budding yeast (red) fits.

Figure S4:
A- Genome browser view of the G4FS found by different algorithms near the promoter. G4FS are represented for G4Hunter with the threshold of 1.5 (green), 1.75 (dark blue) and 2 (light blue). G4FS from QP37 and QP312 are represented in orange and red respectively. Promoter regions of human *KIT*, *BCL2*, *KRAS*, *HRAS*, *SRC* and *TERT* genes are on the pannel top left, top right, center left, center right, bottom left, bottom right respectively.
B- Genome browser view of the G4FS density (#G4FS/kb) found by different algorithms in 200kb region near the promoter represented in S3F. Results for G4Hunter with the threshold of 1.2 (pink), 1.5 (green), 1.75 (dark blue) and 2 (light blue) and G4FS from QP27 (grey), QP312 (red) and QP37 (orange) are represented. *KIT*, *BCL2*, *KRAS*, *HRAS*, *SRC* and *TERT* are on the pannel top left, top right, center left, center right, bottom left, bottom right respectively.
C- Genome browser view of the G4FS found by different algorithm on selected loci from budding yeast genome, fruitfly genome and human rDNA cluster. G4FS are represented for G4Hunter with a threshold of 1.2 (pink), 1.5 (green), 1.75 (dark blue) and 2 (light blue) and for QP37 in orange. The budding yeast loci contains the sequences from Capra *et al*, 2010 and the fruitfly loci are centered arround the ANTP-C and BX-C loci from Hoffmann *et al*, 2015.

Figure S5:
A-Number of hits found by G4Hunter for different thresholds between 1 and 2 for the human genome (hg19) with different window sizes (15: purple, 20: red, 25: blue). The black line indicate the threshold to choose to get the same number of hits as with a threshold of 1.2 for a window of 25: 1.36 for a window of 20 and 1.64 for a window of 15.
B- Overlaps between the hits found by G4Hunter on the human genome with a threshold of 1.2 and a window of 25 (k25_1.2), a threshold of 1.36 and a window of 20 (k20_1.36) and a threshold of 1.64 and a window of 15 (k15_1.64). The numbers within each area indicate the population of each subclass in thousands.
C- Profiles of G4FS arround the transcription start site (TSS) for the UCSC Known Genes list. G4FS list used are, clockwise, G4H1 and 3 settings of Quadparser (QP27, QP312 and QP37). The number on the Y-axis, the fraction of G4FS, represents at the nucleotide level for each position the number of times this nucleotide is found in a G4FS divided by the number of TSS region (39692). The blue and red curves correspond to the G4FS found on the non-coding and coding strands, respectively.

Figure S6:
Profiles of G4FS around the first exon/intron junction for transcript of the UCSC Known Genes list.
G4FS list used are, clockwise, G4H1 and 3 settings of Quadparser (QP27, QP312 and QP37). The number on the y axis, the fraction of G4FS, represents at the nucleotide level for each position the number of times this nucleotide is found in a G4FS divided by the number of junction regions (37466). The blue and red curves correspond to the G4FS found on the non-coding and coding strands respectively.

Supplementary Tables:

Table S1: Reference dataset (excel file)

Table S2: Human mitochondrial genome dataset (excel file)

Table S3: Window size *vs*. Threshold for G4Hunter on the mitochondria genome (excel file)

Table S4: Reference of the genomes used in this study

Species	Reference Genome	Acession number/**R** package reference
Homo sapiens (Human)	hg19, Feb. 2009	BSgenome.Hsapiens.UCSC.hg19
Mus musculus (Mouse)	mm10, Dec. 2011	BSgenome.Mmusculus.UCSC.mm10
Drosophila melanogaster (Fly)	dm3, Apr. 2006	BSgenome.Dmelanogaster.UCSC.dm3
Caenorhabditis elegans (Worm)	ce10, Oct. 2010	BSgenome.Celegans.UCSC.ce10
Dictyostelum discoideum	ddAX4	NC_007087.3, NC_007088.5, NC_007089.4, NC_007090.3, NC_007091.3, NC_007092.3, NC_000895.1, NC_001889.1
Saccharomyces cerevisiae (Budding Yeast)	sacCer3, April 2011	BSgenome.Scerevisiae.UCSC.sacCer3
Schizosaccharomyces pombe (Fission Yeast)	NCBI 2002-03-05	NC_003424.1, NC_003423.1, NC_003421.1,NC_001326.1
Plasmodium falciparum	NCBI 2007-07-24	NC_004325, NC_000910, NC_000521, NC_004318, NC_004326, NC_004327, NC_004328, NC_004329, NC_004330, NC_004314, NC_004315, NC_004316, NC_004331, NC_004317
Escherichia coli	Ecol.NCBI.20080805	BSgenome.Ecoli.NCBI.20080805
Arabidopsis thaliana	TAIR9	BSgenome.Athaliana.TAIR.TAIR9
Macaca mulatta (Rhesus)	rheMac3, Oct. 2010	BSgenome.Mmulatta.UCSC.rheMac3
Pan troglodytes (Chimp)	panTro3, Oct. 2010	BSgenome.Ptroglodytes.UCSC.panTro3
Bos taurus (Cow)	bosTau6, Nov. 2009	BSgenome.Btaurus.UCSC.bosTau6
Sus scrofa (Pig)	susScr3, Aug. 2011	BSgenome.Sscrofa.UCSC.susScr3
Canis lupus familiaris (Dog)	canFam3, Sep. 2011	BSgenome.Cfamiliaris.UCSC.canFam3
Rattus norvegicus (Rat)	rn5, Mar. 2012	BSgenome.Rnorvegicus.UCSC.rn5
Gallus gallus (Chicken)	galGal4, Nov. 2011	BSgenome.Ggallus.UCSC.galGal4
Danio rerio (Zebrafish)	danRer7, Jul. 2010	BSgenome.Drerio.UCSC.danRer7
Apis mellifera (Honey Bee)	apiMel2, Jan. 2005	BSgenome.Amellifera.UCSC.apiMel2
Oryza sativa (Rice)	MSU7	BSgenome.Osativa.MSU.MSU7

Table S5: Number of hits with G4hunter using a window of 25 nucleotide and a threshold indicated in the first column. Note that to calculate the length of the sequenced genome, unattributed bases N have been excluded.

Threshold	H.sapiens	M.mus.	D.mel.	C.elegans	D.discoi.	S.cer.	S. pombe	P. falci.	E.Coli	A.thaliana
1	6939028	6177504	255675	81939	9828	8483	6802	4057	84233	84114
1.25	2890423	2724011	98871	26853	2663	1832	1446	1497	18441	18899
1.5	1436277	1515678	48692	11222	1058	510	391	436	4832	5059
1.75	707106	912630	24281	5249	460	155	108	140	1209	1544
2	339981	569733	12285	2909	249	55	39	47	358	542
Genome (bp)	*2.86E+09*	*2.65E+09*	*1.62E+08*	*1.00E+08*	*3.40E+07*	*1.22E+07*	*1.25E+07*	*2.29E+07*	*6.48E+07*	*1.19E+08*

Table S5A: Reference genomes are respectively hg19(H.s.), mm10(M.m.), dm3(D.m.), ce10(C.e.), ddAX4(D.d.), sacCer3(S.c.), NCB.I20020305(S.p.), NCBI.20070724(P.f.), Ecol.NCBI.20080805(E.c.) and TAIR9(A.t.). Note that the E.c. reference genome contains 13 genomes of different E.Coli strains. Genome (bp) representent the number of base pairs in the genome considered after exclusion of the N.

Threshold	Macaque	Chimpanzee	Cow	Pig	Dog	Rat	Chicken	Zebrafish	Bee	Rice
1	6215253	6585134	6621552	5864845	6468370	6310723	2066700	1539684	287859	854603
1.25	2550419	2722752	2989603	2743767	3106243	2780494	830554	558405	143660	385809
1.5	1223837	1339071	1624638	1511817	1713852	1478994	392312	257457	83169	193640
1.75	574281	654917	864493	842767	955554	834107	194874	123224	47763	95300
2	259771	312863	459403	462330	529573	484128	99909	63453	26135	44648
Genome (bp)	*2.56E+09*	*2.75E+09*	*2.64E+09*	*2.32E+09*	*2.32E+09*	*2.57E+09*	*9.94E+08*	*1.35E+09*	*2.25E+08*	*3.75E+08*

Table S5B: Reference genomes are respectively rheMac3(Mac), panTro3(Chimp), bosTau6(Cow.), susScr3(Pig), canFam3(Dog), rn5(Rat), galGal4(Chicken), danRer7(Zebrafish), apiMel2(Bee) and MSU7(rice). Genome (bp) representent the number of base pairs in the genome considered after exclusion of the N.

Table S6: Fraction of genomic features with at least one G4FS (excel file)

Table S7: Fraction of G4FS found in each genomic feature (excel file)

Table S8: G4FS density in each genomic feature (excel file)

4.4 Algorithm performance analysis (Receivers Operating Characteristic)

Two main criteria are considered when evaluating G4-Hunter: the prediction of true- and false-G-quadruplexes, and the precision of the prediction.

True- and false-G-quadruplexes prediction

The prediction of true and false G-quadruplexes consists on determining how precisely the algorithm can identify G-quadruplexes. This step is essential as it also allows estimating the G-quadruplexes percentage. To achieve this step, we considered the results obtained applying Quadparser and G4Hscore algorithms over the dataset. Then, we calculated the ROC curve (Receiver Operating Characteristic) thanks to the package R: pROC [145, 146]. This curve measures the performance of the algorithm in function of the specificity and the sensibility for different thresholds (Fig. 4.6).

Receiver operating characteristic (ROC)

Receivers operating characteristic (ROC) curves, derived from signal detection theory [147], were first applied for the study of data from psychology [148] and medical fields such as cancer [146]. They are now commonly used in bioinformatics [149], data mining and machine learning, evaluation of biomarker performances and comparing scoring methods. One of the earliest adopters of ROC graphs in machine learning was Spackman (1989), who demonstrated the value of ROC curves in evaluating and comparing algorithms (tom fawcett article). The current popularity of ROC analysis in bioinformatics may be due to the visibly increasing use of machine learning techniques in computational genomics. So, current bioinformatics applications of ROC analysis use concepts and approaches taken from a variety of fields [149].

ROC curves are two-dimensional graphs in which sensitivity is plotted on the Y-axis and specificity is plotted on the X-axis (Fig. 4.7) [150]. Each point on the ROC curve displays relative trade-off between sensitivity (true positives (TP)) and specificity (true negatives (TN)) (see equation 4.1). The determination of an ideal performance becomes difficult as both (specificity- sensitivity) change with each cut-off value [150]. The best cut-off value provides both highest sensitivity and the highest sensibility.

$$\boxed{\begin{array}{l}\textbf{Sensitivity} = \frac{TP}{P} = \frac{TP}{TP+FN} \\ \textbf{Specificity} = \frac{TN}{N} = \frac{TN}{TN+FP} \\ \textbf{Accuracy} = \frac{TP+TN}{P+N}\end{array}} \quad (4.1)$$

Values	Class
1	+
0.98	+
0.84	+
0.71	+
0.69	–
0.69	+
0.64	+
0.62	+
0.56	–
0.5	+
0.39	–
0.37	–
0.36	+
0.32	+
0.32	–
0.29	–
0.28	–
0.25	–
0.11	–
0.00	–
AUC	0.87

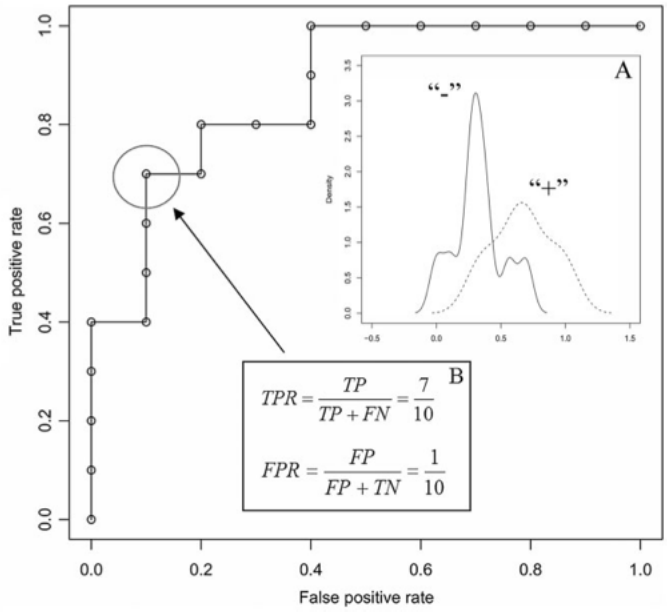


Figure 4.6: **Example of ROC curve.** Twenty tests are classified depending on their score (left). Each score threshold is associated to a specific rate of TP and FP (right, A-B). Adapted from [151]

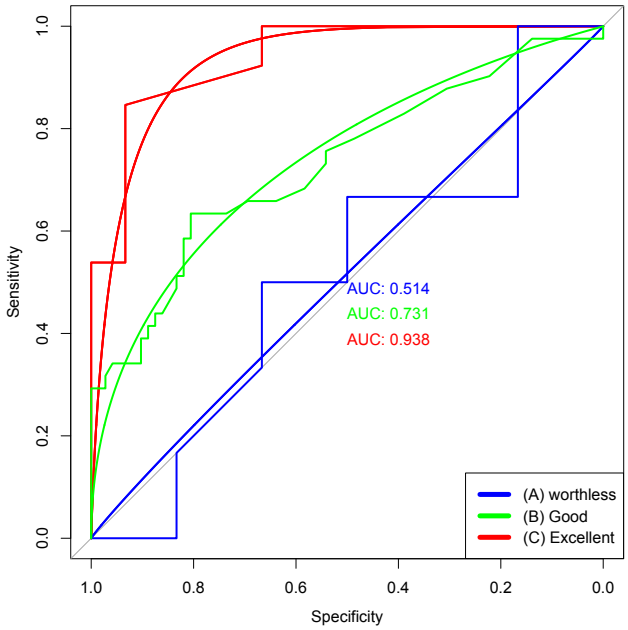


Figure 4.7: **ROC curves with different area under curve's values.** Considering the area under curve, test "C" is better then both "A" and "B" tests, and the curve is close to perfect discrimination. the test "B" had a good validity, where as the test "C" is the worthless.

Where:

- **TP** corresponds to a number of G-quadruplexes sequences in our case, **TN** corresponds to the non-G-quadruplexes sequences, **FP** corresponds the G-quadruplexes that are not detected and the **FN** correspond to the non-G-quadruplexes sequences detected as G4.
- Sensitivity is the proportion of actual positive cases which are correctly identified.
- Specificity is the proportion of actual negative cases which are correctly identified.
- Accuracy is the proportion of the total number of predictions that was correct.

ROC curves Interpretation

When dealing with a dataset that support two labels: positive and negative (Yes, No), we generally assign a class to each of them: one class as a Positive and a class as Negative (see equation 4.2, & Fig.4.6), however some assignments are wrong (False Positive and the False Negative). The ratio between positive and negative events can vary during time. The Receiver operating characteristic curves measure the quality of our test independently of this ratio.
A test is better when its ROC curve approaches the left top corner of the graph (Fig.4.7-red curve). However this ideal can rarely be achieved. In stark contrast if a test has zero discrimination, the proportion of true positives is equal to the false positive's proportion; in this case the ROC curve is the identity curve at 45 degrees (Fig.4.7-blue curve). Most data are between these two extremes (Fig.4.7-green curve). Thereby, ROC curves are interpreted by their area under the ROC curve (AUC) [152], that we can estimate the confidence interval by the [153] methods.

The AUC measures discrimination, thus the ability of the test to correctly classify data. We estimate that a classification is:

- Not satisfactory, if AUC ≤ 0.5,
- Less satisfactory, if $0.5 <$ AUC < 0.7,
- Very satisfactory, if $0.7 <$ AUC < 1,
- Perfect, if AUC $= 1$.

$$\boxed{\begin{aligned} \boldsymbol{P} &= \mathbf{TP} + \mathbf{FN} \\ \boldsymbol{N} &= \mathbf{TN} + \mathbf{FP} \end{aligned}} \tag{4.2}$$

It is also possible to estimate the threshold where this AUC is maximal by the Youden Index. Youden index maximizes the difference between specificity and sensitivity (Fig. 4.6-A). It is more commonly used criterion because this index reflects the intension to maximize the correct classification rate. Easily calculated and located on the ROC curve by finding the highest point on the vertical axis and the furthest to the left on the horizontal axis (Fig. 4.6-B).

Estimation of the prediction precision

The precision of the prediction is compared to the capacity of the algorithm to identify the sequences that form G-quadruplexes. This measure is proposed to evaluate the precision of software depending on the sensitivity and Positive Predictive Value (PPV) (eq 4.3)

$$\textbf{Sensitivity} = \frac{TP}{P} = \frac{Numberofcorrectlypredicted G4}{Numberoftrue G4} = \frac{TP}{TP+FP} = \textbf{PPV} \tag{4.3}$$

4.5 Conclusion

Performing six different *in vitro* experiments on almost 200 potential quadruplex-forming sequences was a challenge. The experimental evaluation of G4-Hunter demonstrated (i) the accuracy of G4-Hunter on a dataset that was not "designed" for G4 formation, (ii) and its ability to predict G4 folding with high sensitivity and specificity. The predictive potential of G4-Hunter makes it a new powerful algorithm to track down quadruplex forming sequences.

In addition, several collaborations yielded to identify more G4 sequences on a large number of genomes. In especial, on going collaboration with Dr. Laurent Lacroix (Toulouse) led to the identification of G4FS with different threshold for 20 genomes including the classical model organisms *Homo sapiens, Mus musculus, Drosophila melanogaster, Caenorhabditis elegans, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Plasmodium falciparum, Escherichia coli* and *Arabidopsis thaliana*. We also collaborate with Dr. Souheila Amor (a form post-doctoral fellow in our lab) to analyze the genome of *Dictyostelium discoideum*, which has a GC poor (22%) genome.

The number of potential G4FS depends on genome size. To facilitate comparisons, we calculated the density in G4FS per thousand of base pairs (kb). The occurrence of G4FS is found to be the highest for vertebrates among the genomes studied, ranking from 2.5±0.1/kb with a threshold of 1 (likely G4FS) to 0.17±0.04/kb with a threshold of 2 (very stable/highly likely G4FS) (Annex, tables 5.9). Unsurprisingly, the human, chimpanzee and macaque genomes have very close properties regarding the number of G4FS. Outside the mammalian class, the densities of G4FS and their dependency to the threshold are quite diverse. The list of G4FS could grow indefinitely, the more genomes are analyzed the more new PG4 will be predicted. The step missing in this work is an *in vivo* analysis in order to fit the usual pathway of predictions. A collaboration with Dr Samir Amrane and Dr Marie-Line Andreola in now in course to complete this work. The results obtained from this collaboration are exposed in the next chapter.

Application to pathogens

Chapter 5

G4s in Viruses, is there a hidden link?

Contents

5.1 Introduction

Human genome sequencing highlighted the presence of many sequences enriched in guanines that can potentially form G-quadruplexes [19, 88]. In parallel, a rising number of exciting reports has suggested roles of G4 structures in several important human microbial pathogens. Several bacteria, protozoa as well as viruses genomes were analyzed to detect new potential quadruplex sequences. Thus, G4 motifs may offer a mean to regulate DNA and or RNA dynamics and viral latency. In this introduction we will summarize the role and function of the known G4s.

5.1.1 Function of G-quadruplexes in different pathogens

Antigenic variation

Antigenic Variation (AV) is one of the most important biological processes in which G4s have been involved in the context of different pathogens. Antigenic variation (AV) is the process by which pathogens express different versions of their surface epitopes in order to evade detection by the host immune system.

G4s are now implicated in AV in several pathogens; the best characterized example is *Neisseria gonorrhoeae*, that causes the human sexually transmitted disease gonorrhea. *N. gonorrhoeae* escape immune inspection through antigenic variation of surface structures such as pili[1]. Pili are mainly composed of pilin proteins, which vary by gene conversion. The antigenic variation uses a G4 to switch the expression of its cell-surface pilin proteins. The major pilus subunit is PilE: antigenic variation occurs following genetic recombination of the PilE expression locus (pilE) with one or more silent pili loci (pilS).

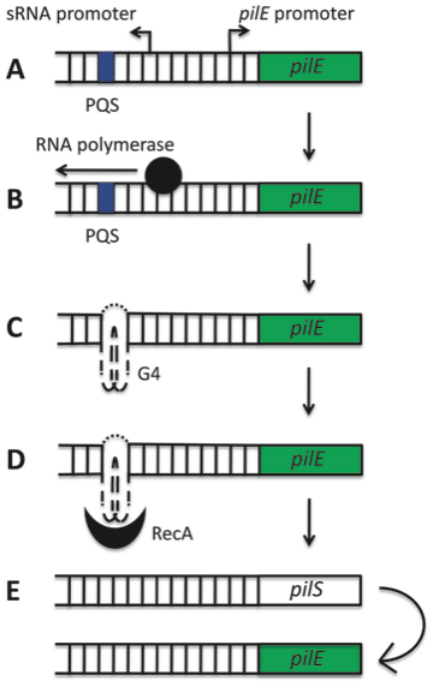


Figure 5.1: **Schematic model of the role of the pilE G-quadruplex (G4) in *N. gonorrhoeae* pilin antigenic variation**. A putative quadruplex sequence (PQS) and a small RNA (sRNA) promoter are located upstream of the pilE locus (A). The initiation of transcription from the sRNA promoter (B) provides the single-stranded conditions required for G4 formation (C). The pilE G4 recruits RecA (D) and potentially other recombination factors, which stimulates non-reciprocal recombination between a pilS locus and the pilE locus (E). Adapted from [154, 78].

[1]A pilus (Latin for 'hair'; plural : pili) is a hairlike appendage found on the surface of many bacteria.

The recombination initiation site has been mapped to a 16 base pairs G-rich segment ($G_3TG_3TTG_3TG_3$) that forms a parallel intramolecular G4 structure *in vitro* and is located upstream of the pilE locus [78, 155, 154]. The G-quadruplex-forming pilE sequence is only present at that location in the gonococcal genome (upstream of pilE) (Fig 5.1). The requirement of the RecA helicase for pili AV suggest that the monomeric pilE G-quadruplex recruits RecA to promote recombination between pilE and a pilS copy [155].

Another example is the Lyme spirochetes *Borrelia spp.* that also evades the host immune response through AV. Walia R. *et al*, have showed that G4 DNA formed by sequences within the B31 vlsE locus occur at very high densities in three *B. burgdorferi* strains as well as in *B. afzelii* and *B. garinii*. This suggests that G4 DNA may be a mediator of recombination switching at the vlsE locus [156].

Outside telomeres, PQS are rare in the AT-rich genome of the malaria parasite *Plasmodium falciparum*. Interestingly, 16 non-telomeric PQS are located 1612 to 1707 bp upstream of VAR family genes [157]. The var gene family encodes PfEMP1, the parasite's major variant antigen expressed at the surface of infected erythrocytes, which plays a key role in malaria pathogenesis and immune evasion. The predominance of G4 motifs in var gene regulatory regions suggests that they may play roles in var gene recombination and/or switching. Therefore, it appears that *Plasmodium falciparum* could encode a helicase that may have the capacity to regulate G-quadruplex structures that could be involved in var gene regulation [157].

Translation regulation

Epstein-Barr virus (EBV) is a herpes family virus that can induce infectious mononucleosis and cancers. The virus is composed of a 172 kb double-helix DNA that circularizes upon entry into the nucleus and becomes a viral episome[2]. Latent infection by EBV requires both replication and maintenance of the viral genome. Epstein-Barr nuclear antigen 1 (EBNA1) was the first EBV protein detected the most widely studied. EBNA1 is required for the persistence of EBV genomes due to its contributions to both the replication and mitotic segregation of EBV episomes. The level of EBNA1 synthesized is tightly controlled; it is sufficiently high to maintain viral infection but sufficiently low to avoid immune recognition by the host's virus-specific T cells. The replication of EBNA1 depends on linking region 1 and 2 (LR1 and LR2).

LR1 and LR2 have RNA-binding activity, with a strong preference for G-rich RNA. Destabilization of the G4-forming sequence increases the translation rate and, consequently, promotes antigen presentation. On the other hand, stabilization of the G4 with a G4 ligand (pyridostatin) decreases EBNA1 synthesis and allows immune evasion. These findings suggest that the translational regulation mode may be more general among proteins that self-regulate synthesis. In addition, alternative therapeutic strategies focused on targeting RNA structures within viral ORFs could be developed, in order to interfere with the virus cycle as well as to promote antigen presentation and to stimulate the host immune response [158].

[2] DNA segment that can exist and replicate either autonomously in the cytoplasm or as part of a chromosome.

Control of the host cell's immune response.

The SARS coronavirus (SARS-CoV) is more pathogenic for humans than any other coronavirus. Therefore, the presence of the so-called SARS-unique domain, encoded by the SARS-CoV genome, which is absent in other coronaviruses is interesting, because it may be responsible for the virulence. G-stretches are found in the 39-non-translated regions of mRNAs coding for certain host-cell proteins involved in apoptosis or signal transduction and have been shown to bind to SARS-unique domain (SUD) *in vitro*. Their inhibitory effect on viral replication is being studied with encouraging results. Therefore, SUD may be involved in binding to viral or host-cell RNA and thereby regulate viral replication or controlling the host cell's response to the viral infection [159].

Transcription regulation

The polyomavirus simian virus 40 (SV40) is a known oncogenic DNA virus, which induces primary brain and bone cancers, malignant mesothelioma and lymphomas in laboratory animals [160]. The double-stranded DNA genome encodes for six proteins and includes non-coding regulatory region (NCRR). Notably, six GC boxes (GGGCGG) are present in this region, which can form a quadruplex structure. Replication of the simian virus 40 (SV40) genome requires the large T-antigen (T-ag), a multifunctional protein, which together with host cell factors, initiates viral replication. Interestingly, T-ag can unwind G4 DNA structures and thus might plays a crucial role in regulating replication as well as early and late transcription. Interestingly Perylene diimide derivatives (PDI) stabilize G4 structures and inhibit both the G4 and T-ag duplex DNA helicase activities. This introduces new insights into the link between helicases and tumorigenesis or other human genetic diseases [161].

5.1.2 G-quadruplexes as therapeutic targets

Herpes Simplex Virus-1 genome has a 68% GC composition. G-rich sequences are mainly located in repeats of the HSV-1 genome and can fold in stable G-quadruplex structures. Treatment with the G-quadruplex ligand BRACO-19 greatly stabilizes these sequences resulting in a decrease in the number of infectious viral particles and reduction of late viral transcripts. Taq polymerase processing at the HSV-1 genome, specifically affecting viral DNA replication at G-quadruplex regions in the presence of BRACO19. This work indicates the possibility to block viral DNA replication by G-quadruplex-ligands and therefore provides a proof of concept for the use of G-quadruplex ligands as new therapeutic options [162].

5.1.3 G-quadruplex-forming oligonucleotides with antiviral activity

Aptamers are nucleic acid sequences that can adopt a 3D structure and specifically recognize a given target. These structured nucleic acids generally originated from *in vitro* selection. This is a method of *in vitro* selection from combinatorial libraries of synthetic oligonucleotides

containing millions of different sequences. G-quadruplex forming aptamers present interesting therapeutic potential [163]. The most promising to date is AS1411 (AGRO100), a 26 bp anti-proliferative agents targeting nucleolin [164][3].
AS1411 has recently completed phase II clinical trials as an anticancer drug with low toxicity, highlighting the high therapeutic potential of G4s. It is interesting to note that during the last fifteen years many aptamers adopting G4 structures were selected to target different viral proteins.

In the case of the Hepatitis A virus (HAV), the genome encodes a single polyprotein, which includes the 3C protease. This protease plays a crucial role in the viral life cycle. It cleaves the polyprotein into several proteins. Additionally, it binds to regulatory structural elements at the 5' UTR, thereby controlling viral genome synthesis. The 3C proteolytic activity has been investigated in order to develop anti-viral drugs. Recently, *in vitro* studies showed that a hexanucleotide (GGGGGT) sequence, that forms a tetramolecular G4, binds specifically to the 3C protease of hepatitis A virus. The binding of the G4 could interfere with the protein function. Many experiments strongly indicate that this hexanucleotide (G5T) acts as an inhibitor of HAV-specific gene expression, presumably via its inhibitory effects on 3C protease. It is suggested that further modifications of such ligand may lead to novel inhibitors of virus replication *in vivo* [165].

Another example is the influenza A virus (IAV). Non-structural protein 1 (NS1) of the influenza A virus (IAV) inhibits the host's innate immune response by suppressing the induction of interferons (IFNs). A G4-forming aptamer was selected for it high affinity for the NS1 RNA binding domain. It has the ability to induce INF-β by suppressing the function of NS1. These results indicate that the selected G4 aptamer has strong potential to be further developed as a therapeutic agent against IAV [166, 167].

Finally, several G4 forming aptamers were against HIV-1 viral proteins such as the integrase [168], RT [169, 170], gp120 [171] Rev [172] and the viral Tat cofactor. Some of these aptamers can prevent the entry of the virus into the cell by interacting with the viral protein gp120. Concerning integrase the 93del aptamer potentially binds to the catalytic pocket of the enzyme [173]. This G4 strongly inhibits the infectivity of HIV-1 with an IC_{50} of 25 nM. A functional study of the effects of this aptamer on the infectivity of HIV-1 *in vivo* conditions demonstrated that the fusion, transcription and integration of the provirus are inhibited by 93del [174].

5.1.4 Objectives

These studies show that certain viral proteins recognize very specifically and with high affinity the G4 structures. In the case of HIV-1 the 3 most important viral enzymes are able to bind to G4 structures, Surprisingly, at the beginning of this project almost no G4 motifs were identified in the HIV-1 genome. In this context, we decided to first search for conserved G4 motifs within HIV-1 genome and later extend this approach to Ebola and Marburg viruses.

[3]note that the word "aptamer" may be abusive for AS1411, as it was not isolated by a Selex process

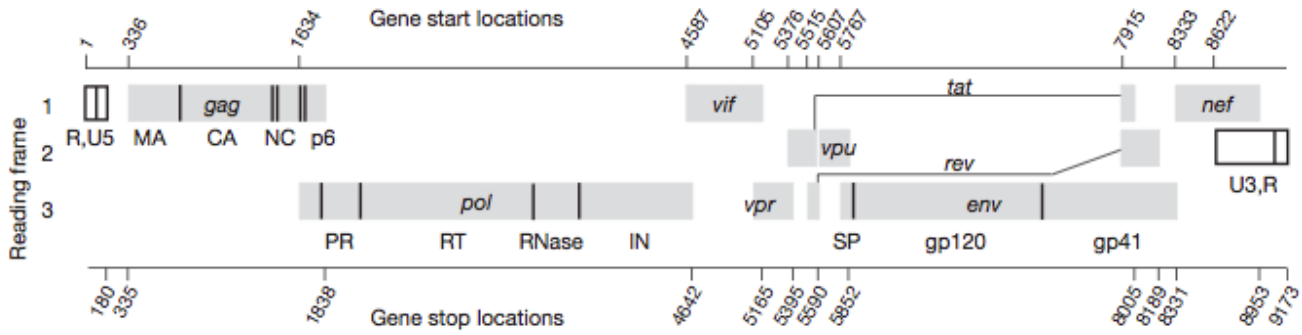


Figure 5.2: **HIV-1 RNA genome organization** Protein coding region are shown as grey boxes; polyprotein- domain junctions are depicted as solid vertical lines.CA, capsid; IN, integrase; MA, matrix; NC, nucleocapsid; PR, protease; RT reverse transcriptase; SP, signal peptide. Adapted from [175]

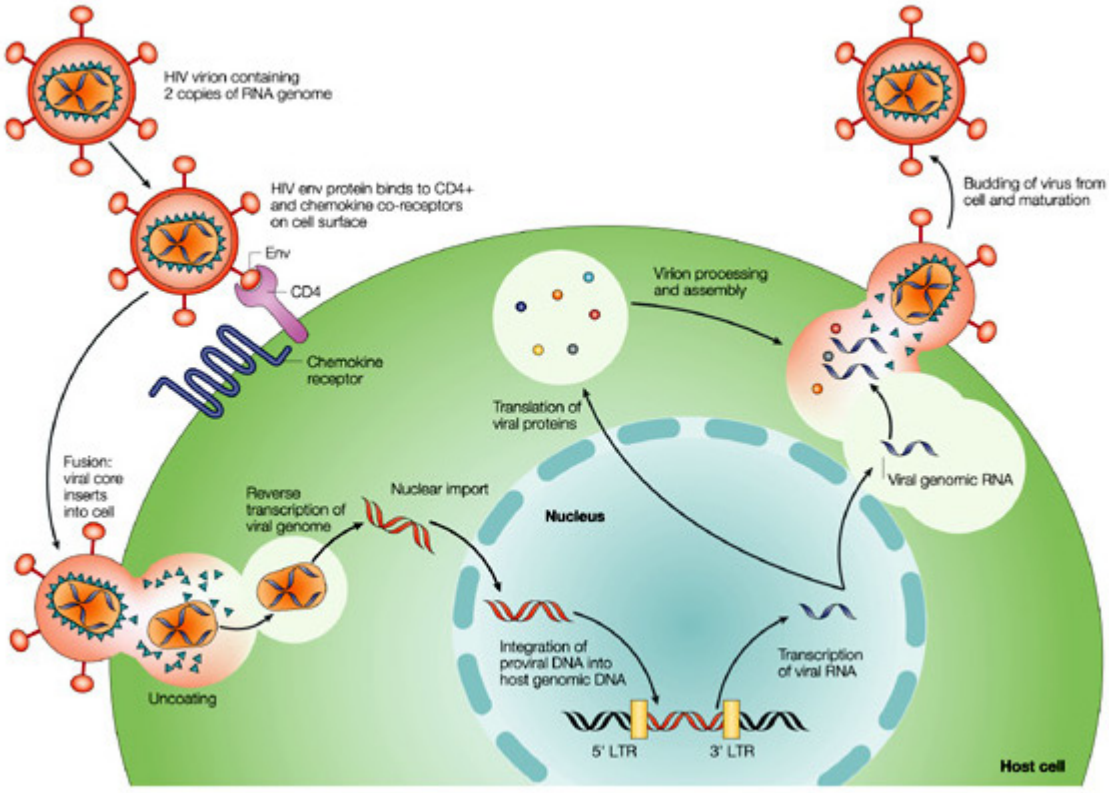


Figure 5.3: **HIV-1 replication cycle.** Adapted from [176].

5.2 Hunting new G-quadruplexes in HIV

5.2.1 Human Immunodeficiency Virus

An estimated 39 million people live with HIV-1, while about 25 million have died already. Heterosexual transmission remains the main mode of transmission and accounts for about 85% of all HIV-1 infections. Women make up about 42% of those infected worldwide and HIV-1 infection in women has additional implications for mother-to-child transmission [177]. Even without a vaccine, development of highly active antiretroviral therapies has allowed people to live with HIV as a chronic disease [178]. However, viruses within T cells remain fully capable of replicating and infecting other cells if the drug pressure is removed or when resistance emerges. Thus, new drugs must be developed to overcome this issue [179].

HIV-1 is an RNA virus in the *Lentivirus* genus and is part of the *Retroviridae* family. Lentiviruses are single-stranded, positive-sense, enveloped RNA viruses. Two type of HIV have been characterized: HIV-1 and HIV-2. Based on their genetic organization, HIV-1 viruses are divided into three groups (major M, N and O group). These HIV-1 groups and HIV-2 probably result from distinct cross-species transmission events [178]. The worldwide spread of HIV-1 indicates that the virus effectively counteracts innate immunity [180].

The HIV-1 virion has a roughly spherical shape of 110 nm diameter. It consists of an envelop that surrounds a conical capsid. The envelop, is composed of a host-cell derived lipid membrane and is coated in the interior by the matrix proteins (MA). On the envelop two glycoproteins are anchored: the trimer gp120 surface protein and the gp141 transmembrane protein complex. Moreover, the capsid contains the viral enzymes integrase and reverse transcriptase (RT) and the four accessory proteins.

The 9.18 kb HIV-1 RNA genome (Figure 5.2, Table 5.1) consists in (i) three primary genes (gag, pol and env), (ii) two regulatory genes (tat and rev) and (iii) four accessory genes (vif, vpr, vpu, nef). Gag, Pol and Env are the prototypical retroviral proteins [181]: gag encodes for capsid proteins (MA, CA, NC and p6), pol encodes for the three viral enzymes protease (PR), reverse transcriptase (RT) and integrase (IN) and env encodes for the two viral glycoprotein gp120 and gp141. Gag, Pol and Env products are essential for viral replication. In addition, two regulatory proteins Tat and Rev are also essential. On the contrary, the four so-called accessory proteins Nef, Vif, Vpr and Vpu are not essential for the *in vitro* replication, but can have consequences on viral life cycle, altering replication or disease progression [182].

HIV-1 particles contain two copies of a single-stranded RNA. The viral reverse transcriptase (RT) converts the RNA into double-stranded DNA as a part of viral replication. The resulting viral DNA is then imported into the nucleus and the insertion into the cellular DNA is catalyzed by the virally encoded integrase (IN). Once integrated, transcription regulated by the viral promoter at the 5' long terminal repeat generates mRNAs that code for various viral proteins and genomic RNA. Alternatively, the provirus may become latent, which allows the virus and its host cell to avoid detection by the immune system. The HIV-1 life cycle is complex (Fig 5.3) and its duration and outcome depend on the target cell type and cell activation.

Table 5.1: HIV-1 genome organization and products

Start[a]	**End**[b]	**Name**[c]	**Start**[d]	**End**[e]	**Gene**	**locus_tag**	**Products**	
357	388	gag1	336	1838	gag	HIV1gp2	Pr55(Gag)	matrix/ capsid/p2/ nucleocapsid/P1/p6
1928	1964	pol2	336	4642	gag-pol	HIV1gp1	Gag-Pol/ polyprotein	Pol aspartic peptidase p66 subunit reverse transcriptase p51 subunit integrase
4194	4219	pol3						
4324	4377	pol4						
4499	4527	pol5						
5245	5280	vpr	5105	5396	vpr	HIV1gp4	Vpr/p15	artificial frameshift
5982	6022	env7	5377	7970	tat	HIV1gp5	p14	
			5516	8199	rev	HIV1gp6	p19	
7535	7575	env8	5771	8341	env	HIV1gp8	Envelope surface glycoprotein gp160 precursor	
8111	8150	env9	5516	8199	rev	HIV1gp6	p19	
			5771	8341	env	HIV1gp8	Envelope surface glycoprotein gp160, precursor Envelope surface glycoprotein gp120 Envelope transmembrane domain	hypothetical protein
8608	8649	nef10	8343	8963	nef	HIV1gp9	Nef/p27	
			8631	9085	3'UTR			

[a]Start of PQS
[b]End of PQS
[c]Arbitrary name of PQS according to their position in the genome
[d]Start of the gene
[e]End of the gene

Table 5.2: PQS obtained by G4-Hunter score calculation of the NC_ 001802 HIV-1 sequence.

Start	End	Sequence (5' -3')	Length	Score
258	283	GCGCACGGCAAGAGGCGAGGGGCGG	25	1.04
260	297	GCACGGCAAGAGGCGAGGGGCGGCGACTGGTGAGTAC	37	0.81
357	388	TAAGCGGGGGAGAATTAGATCGATGGGAAAA	31	1.0
393	418	GGTTAAGGCCAGGGGGAAAGAAAAA	25	1.0
870	895	CCACCCCACAAGATTTAAACACCAT	25	-1.0
1686	1712	GCAGACCAGAGCCAACAGCCCCACCA	26	-0.96
1690	1715	ACCAGAGCCAACAGCCCCACCAGAA	25	-1.0
1833	1868	AATAAAGATAGGGGGGCAACTAAAGGAAGCTCTAT	35	0.74
1928	1964	CAAAAATGATAGGGGGAATTGGAGGTTTTATCAAAG	36	0.78
2868	2902	GAAGTTAGTGGGGAAATTGAATTGGGCAAGTCAG	34	0.85
3415	3443	TATGTAGATGGGGCAGCTAACAGGGAGA	28	0.93
4162	4188	TGTTGGTGGGCGGGAATCAAGCAGGA	26	0.96
4194	4219	AATTCCCTACAATCCCCAAAGTCAA	25	-1.04
4324	4377	AATTTTAAAAGAAAAGGGGGGATTGGGGGGTACAGTGCAGGGGAAAGAATAGT	53	1.26
4499	4527	TCTGGAAAGGTGAAGGGGCAGTAGTAAT	28	0.89
4671	4708	TGTATGTTTCAGGGAAAGCTAGGGGATGGTTTTATAG	37	0.84
5245	5280	GAAACTTATGGGGATACTTGGGCAGGAGTGGAAGC	35	0.91
5789	5833	ATCAGCACTTGTGGAGATGGGGGTGGAGATGGGGCACCATGCTC	44	0.89
5982	6022	ACATGCCTGTGTACCCACAGACCCCAACCCACAAGAAGTA	40	-0.88
6404	6429	CAATTCCCATACATTATTGTGCCCC	25	-1.0
7535	7575	ATCAACAGCTCCTGGGGATTTGGGGTTGCTCTGGAAAACT	40	0.7
7913	7946	TATCGTTTCAGACCCACCTCCCAACCCCGAGGG	33	-0.85
7925	7951	CCCACCTCCCAACCCCGAGGGGACCC	26	-1.15
8111	8150	TTGTGGAACTTCTGGGACGCAGGGGGTGGGAAGCCCTCA	39	0.79
8230	8266	GCAGTAGCTGAGGGGACAGATAGGGTTATAGAAGTA	36	0.81
8517	8544	AAGCACAAGAGGAGGAGGAGGTGGGTT	27	0.93
8521	8546	ACAAGAGGAGGAGGAGGTGGGTTTT	25	1.0
8608	8649	CTTTTTAAAAGAAAAGGGGGGACTGGAAGGGCTAATTCACT	41	0.8
8990	9017	GCTGGGGACTTTCCAGGGAGGCGTGGC	27	1.0
9001	9045	TCCAGGGAGGCGTGGCCTGGGCGGGACTGGGGAGTGGCGAGCCC	44	0.86

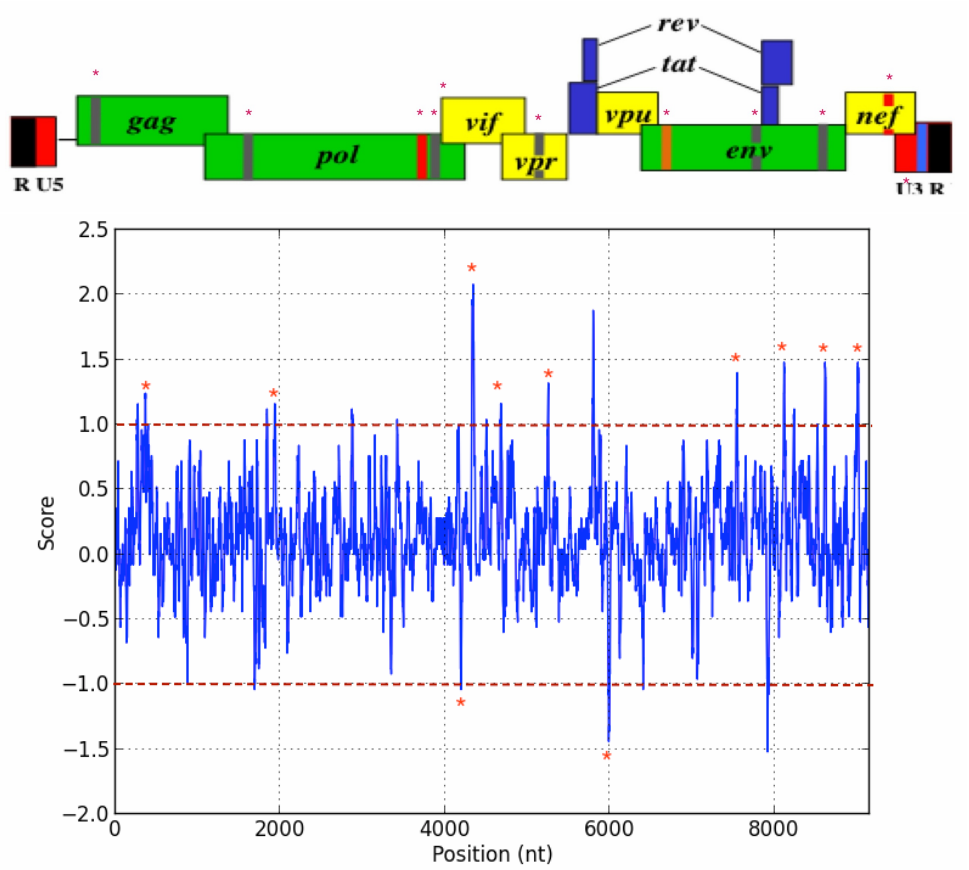


Figure 5.4: **Score calculation of the NC_ 001802 HIV-1 sequence.** All the conserved PQS (score $\geq|$ 1 $|$) are represented by stars on both the graphical representation of the HIV genome (top) or on the top of each peak (bottom).

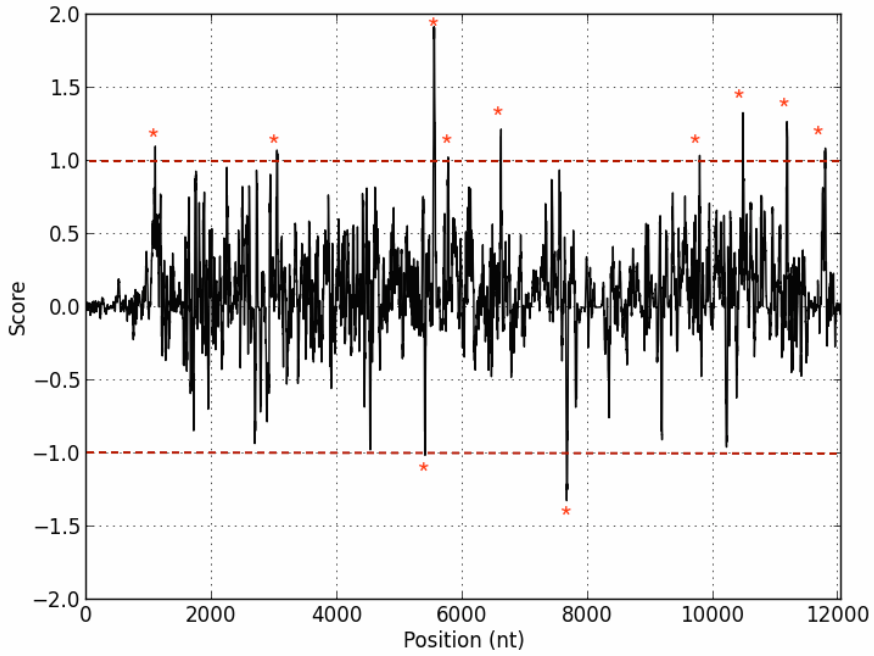


Figure 5.5: **Score calculation for 2177 aligned HIV-1 sequences.** The difference in position is due to the alignment process.

In the early steps, access to cells occurs without causing immediate lethal damages but the entry process can stimulate intracellular signal cascades, which in turn might facilitate viral replication [177].

Current drug treatments have revolutionized the treatment of HIV (AIDS). Typically, they consist of various combinations of compounds targeting the viral proteins reverse transcriptase, protease and glycoprotein gp120. However, a number of problems with current therapies limit their usefulness. First, the cost of the drugs constitutes a significant burden to individuals and governments worldwide and virtually eliminates their availability in developing countries. Additional problems include inconvenient and complicated medication schedules, the lack of patient compliance, side-effects associated with the drugs and ominously, the development of drug-resistance. For these reasons, alternative treatment strategies for HIV infection are being investigated [183].

In this context, a collaboration with Dr. S. Amrane in the lab & Dr. ML. Andreola (CNRS, Bordeaux) has been initiated (i) to perform a comprehensive analysis of G4 sequence conservation in the HIV genomes and assess the possibility to find new motifs, (ii) to test the ability of HIV viral proteins to recognize those G4 localized in the same HIV genome (iii) and to find out what are the G4 ligand effects on the HIV activity. The next part of this chapter will focus on presenting the work that we have been doing and the different results obtained.

5.2.2 G-quadruplexes and HIV: the hidden link

G4-Hunter analysis

At the beginning of the project (2013) very few data were available regarding the presence of G-quadruplexes in the HIV-1 genome. It was previously suggested that two G-quadruplexes are present in the gag and pol genes. They could be involved in some recombination events [188, 189, 190, 191]. However a complete screening of G4 motifs in the HIV-1 genome had never been performed at that time. We therefore decided to apply G4-Hunter to predict G4 formation on the HIV genome. We first searched for G4 motifs within a single specific HIV-1 genome isolate (GeneBank Accession number: NC_001802) using G4-Hunter and Quadparser. 30 potential G4 candidates were predicted with G4-Hunter using a threshold of 1. We found 22 sequences on the coding strand and 8 on the noncoding strand (Fig 5.4 & Table 5.2). Using Quadparser only 9 sequences were identified. However, it is well known that the HIV-1 RNA genome is rapidly evolving, which gives it an important structural variability allowing it to escape the immune response of the infected organism. The low fidelity of the reverse transcriptase and the many genetic recombination events occurring between the two viral RNAs are the drivers of this trend [191].

To investigate the conservation of these motifs, we searched those potential G4 sequences in 2177 HIV genomes. We used the HIV genome database (http://www.hiv.lanl.gov/) in order to extract the genomic data. The database contains information on HIV genetic sequences, drug

Table 5.3: Conserved HIV-1 prone motifs and the *in vitro* conclusion.

Start	Start[a]	Name	Reference	Sequence	Length	Score	Conservation[b]	Conclusion 4 µM	Tm
357	1095	GAG	[184, 185]	GGGGGAAAATTAGATGCATGGGAGAAAATTCGGTTAAGGCCAGGGGGA	48	1.13	61	Yes	27
1928	3035	pol2		AAATGATAGGGGGAATTGGAGGTTTTATCAAAGTA	35	0.83	976	Not G4	
4194	5402	pol3		CCTTGACTTTGGGGATTGTAGGGGATC	27	1.04	950	Not G4	
4324	5535	C-ppt	[186]	AGAAAAGGGGGGATTGGGGGGTACAGTGCAGGGG	37	1.86		G4*+ [c]	35/42
4499	5769	pol5		TGGAAAGGTGAAGGGGCAGTAGTAATAC	28	0.89	1560/1852	no	
5245	6609	vpr1		TGGGGATACTTGGGCAGGAGTGGAAGC	27	1.19	252	G4	27
		vpr2		TGGGGATACTTGGGAAGGGGTTGAAG	26	1.62	223	G4*+	11/_30
5982	7661	env7		TGTGGGTTGGGGTCTGTGGG			1062	G4	20
7535	9784	env8		AGGAATTTGGGGCTGCTCTGGA	22	1.0	332	G4?	30
				GTCTCAAAGGGTTGAGGAGGGGGTGGGAAGGCCTC	35	1.17		G4*+	43/nd
8111	10471	env9		GGGACTACAGAGGGGGTGGG	20	1.85	183	G4*+	63/78
				GGGACTGAGACTGGGGTGGG	20	1.7	196	G4*+	45/56
8608	11175	nef10 (PPT)	[187]	GAAAAGGGGGGACTGGAAGGGCT	23	1.57	574	Not G4	

[a]Start on the aligned genome

[b]Number of isolates containing this exact sequence

[c] +: Stable quadruplex with Tm $> 37°C$ & *: Hysteresis in the melting process.

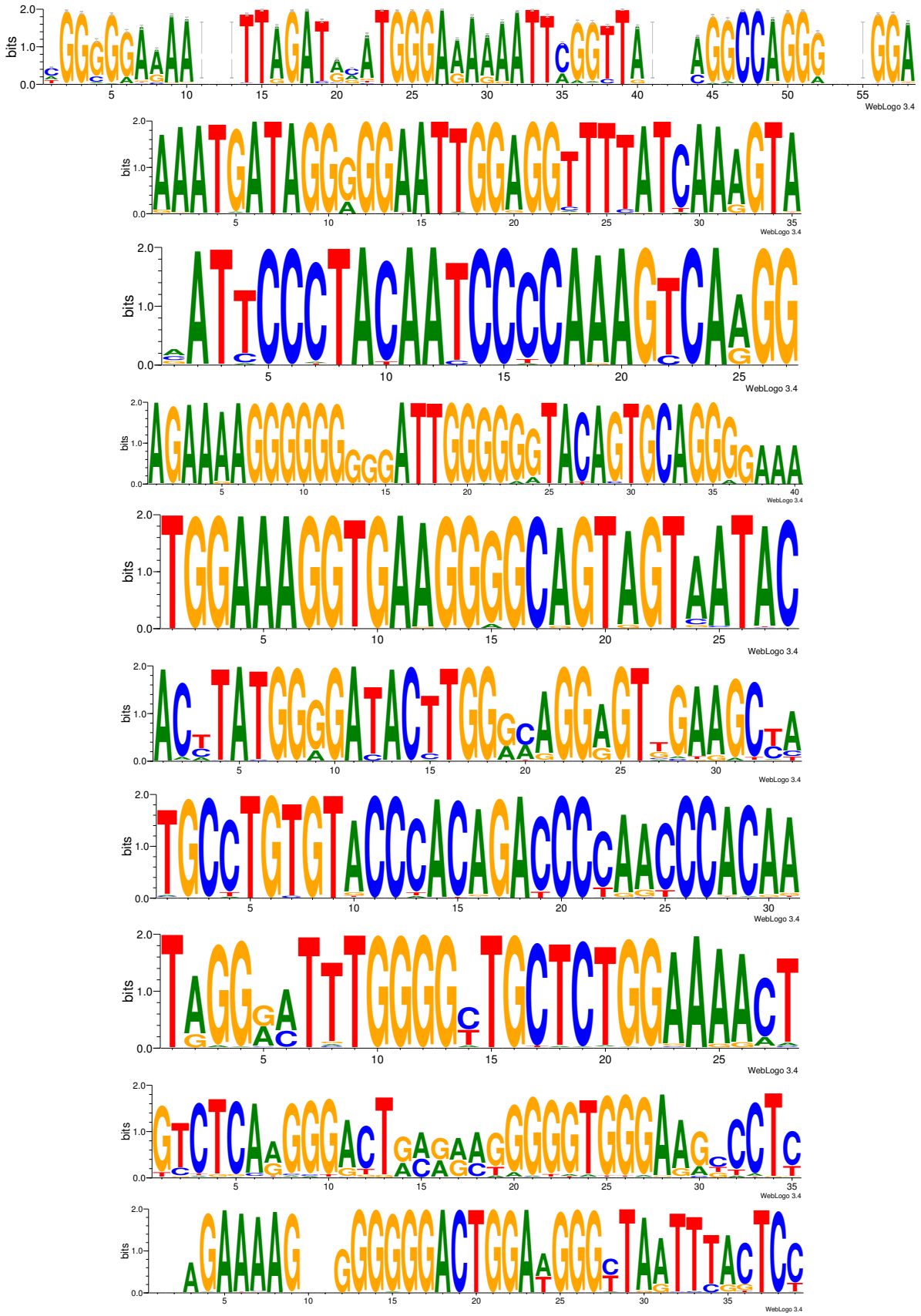


Figure 5.6: **LOGO representation of the different HIV-1 PQS.** From the top to the bottom: GAG, pol2, pol3, C-ppt, pol5, vpr, env7, env8, env9, nef.

resistance-associated mutations and vaccine trials. The website gives access to a number of tools that can be used to analyze these data such as web alignments. The multiple alignments was generated for 2177 HIV-genomes and downloaded in a Fasta file format. In this section we present the steps used to extract conserved G-quadruplexes and their potential role.

Conserved sequences score calculation

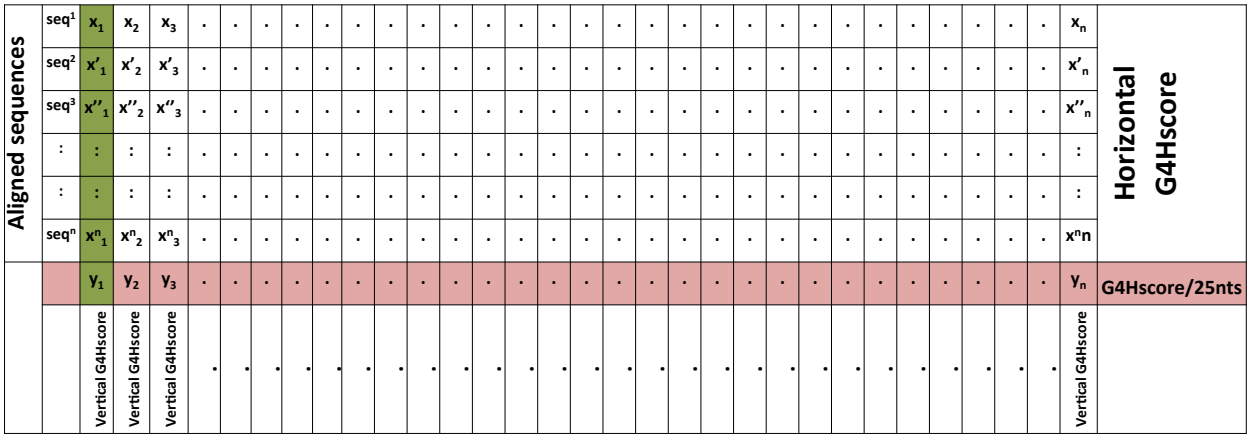
All HIV-1 subtypes are represented in this alignment which contains genomes from various parts of the world. Sequence alignment is a process of arranging the sequences (DNA, RNA, or protein) to identify similar regions that might be a consequence of functional, structural, or evolutionary relationships between the sequences. Aligned nucleotide sequences are represented as rows within a matrix. Generally, gaps symbolized by dashes "-" are inserted between the residues so that identical or similar characters (bases or amino acids) are aligned in successive columns.

Our goal was to extract conserved G4 sequences within the 2177 genomes. This is achieved by calculating the score using G4-Hunter. The principle of G4-Hunter was applied for each sequences (see chapter 3). Briefly, a score of 1/-1 was attributed to each single G/C base, a score of 2/-2 for each GG/CC, 3/-3 for each triplet GGG/-CCC and 4/-4 for a quadruplet or more G/C. The presence of a gap "-" in a sequence was considered as neutral or indifferent bases (like the adenine or thymine) because they are present only to optimize the alignment. Thus, the score is set to 0 for each A, T and dashes "-".

When dealing with multiple aligned sequences, the G4score calculation had to be modified. We considered the alignments files as a matrix. The numbers of matrix lines are represented by the numbers of sequences aligned and the column are represented by the numbers of bases of each sequences (Fig 5.7). To calculate the score of this matrix a new line is added at the end (Fig 5.7-red line). This line came in order to create a new scored nucleic acid sequence with the following characteristics:

- Same length of the all aligned sequences;
- Each nucleotides score has a vertical G4Vscore (Y), which is the arithmetic mean of all nucleotides score of the same column (position).

Thereby, as described earlier we converted each base into a number according to G4-hunter rule (Horizontal G4Hscore). The average score or Horizontal score of each nucleotide is then calculated. This nucleotide score is the basis for G4Hscore, stocked in a list and represents the last line of the matrix. The last step is then to calculate the G4Hscore for each window of 25 nucleotides. Regions in which the absolute value of G4hunter score is above the threshold "1" are extracted and considered as a conserved PQS.



$$y_1=(x_1+x'_1+x''_1+.....+x^n_1)/N \quad \text{(a)}$$

x : G4hunter base score
N: number of aligned sequences

Figure 5.7: **Matrix organization of an aligned sequences and principal of score calculation for aligned sequences.** Each sequence base has a score (x_1, x_2 ..... x_n) according to G4-Hunter rule (which represent G4Hscore or Horizontal score). For each base a vertical score is calculated according the the formula (a). The final score is calculated from the last line of the matrix (red line)

Using this approach, we were able to identify 11 conserved PQS. These sequences represent sequences with G4Hscore $\geq|$ 1 | (Fig 5.5 11 starts). The name of the sequences was given according to the name of the major gene where are located (i.e. env7: is the seventh PQS and it is present in the vpr gene (Table5.1). Nine sequences (GAG, pol2, C-ppt, pol5, vpr1& 2, env8, env9-1 & 2, nef10 (PPT) and the promoter PRO1) are present on the coding (+) strand and 2 sequences (pol3, env7) on the noncoding (-) strand. In order to get the precise description of the relative frequency of bases at their relative position, sequences logo were created for all the 11 PQS by using the weblogo website (Fig 5.6, and Fig 5.9).

Despite the high genetic variability of HIV-1, it is interesting to note that the sequences are highly conserved across all sub-species of HIV-1 as shown in the logo representation. It is also showing the conservation of each base of the alignment. In some cases, the blocks of G's are conserved, with slight differences in the number of G's from the blocks (C-ppt, pol5, env8, env9). The linkers (loops) are sometimes heterogeneous in length and base composition (vpr, GAG, env9); however, despite this sequence heterogeneity, the ability to form G4s is conserved in each single sequence.

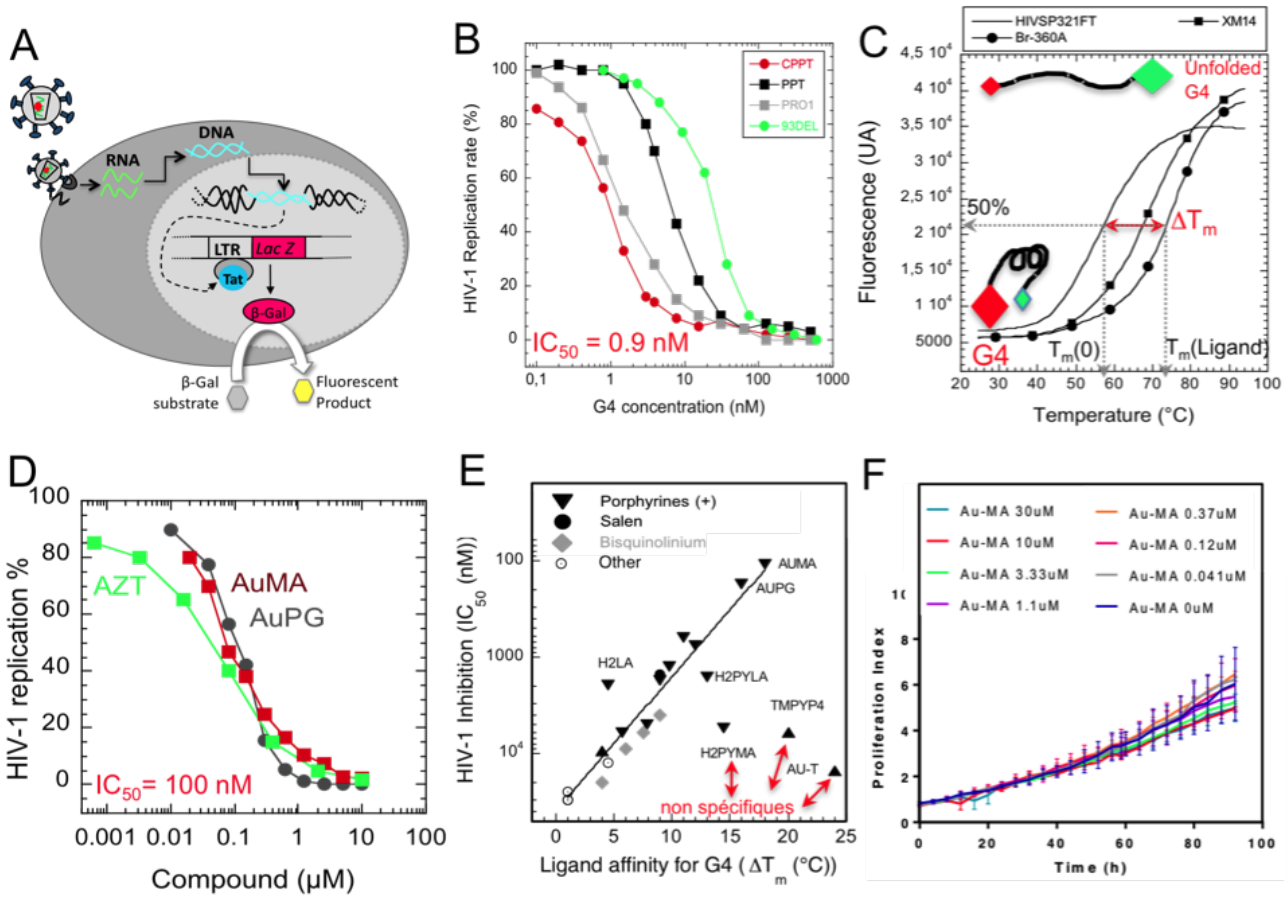


Figure 5.8: ***In vivo* validation of G-quadruplex formation** (A). HeLa P4 cell system for studying the replication of HIV-140. The HeLa P4 cells contain a gene lacZ encoding beta-galactosidase under the control of the viral LTR and whose transcription is activated by the viral protein Tat. The cells are incubated with the ligand or the viral G4 for 30 minutes and then infected with HIV-1. After 24 hours the activity of beta-galactosidase, which is proportional to the infectivity of the virus is then measured on a fluorescence plate reader. (B). Effects of dose of 3 viral G4 (decoy) and 93del aptamer. (C). Principle of the stabilization test by FRET: the transfer of fluorescence energy between fluorescein and tetramethylrhodamine is possible when two fluorophores are close in the folded state. (D). Examples of dose effects observed for the 2 most powerful G4 ligands and AZT. (E). Correlation between the apparent affinity of the ligand for the G4 (expressed in stabilization of the G4 in ΔTm$°C$) and its inhibition activity on HIV. (F). Measurement of the cell proliferation of HeLa cells as a function of time and the concentration of Au-MA G4 ligand.

In vitro validation of G-quadruplex formation

In order to confirm the G4 folding of those conserved sequences, we tested the ability of those sequences to fold into G4 structure by performing different *in vitro* assays described earlier.[4] The results obtained are summarized in the Table 5.3.

All the sequences were tested under 100 mM KCl and 4 µM strand concentration. In this group, three sequences were not able to fold into G4 structure (pol2, pol3 and pol 5) under these conditions, whereas the remaining sequences were shown to form G4 structures, as demonstrated by the IDS (with or without salt at $20°C$) or by thermal difference (TDS) (data not shown). The temperature of half dissociation was determined for all those sequences (Table 5.3). Stable sequences are characterized by a Tm $> 37°C$ (C-ppt, env9-1 and env9-2). In addition, the circular dichroism spectrum recorded for those sequences indicate their parallel or antiparallel topology.

In vivo validation of G-quadruplex formation

In collaboration with Dr. Marie-Line Andréola, we opted for two strategies to explore the potential role of all these G4s in the HIV viral cycle:

The decoy strategy In this first strategy we wanted to know if the G4s contained in the genome of HIV-1 were able to inhibit the virus like the 93del and T30177 aptamers. We tested the ability of the G4 identified in the genome of HIV-1 to inhibit the replication of the virus on HeLa P4 cells infected by HIV (Fig 5.8-A). In the presence of synthetic viral G4s from the HIV genome, we observed extremely powerful inhibitions.
C-ppt, 3'PPT and the G4 promoter (PRO1) sequences, pre-incubated in 100 mM potassium solution to promote G4 formation, are able to inhibit HIV with IC_{50} of about 1 to 10 nM (Fig 5.8-B). These effects are much higher than those obtained in the presence of 93del (IC_{50} = 25 nM), which is one of the best anti-HIV oligonucleotides described in the literature. Interestingly, cytotoxicity tests after 24 hours on HeLa P4 cells showed no cytotoxic effect of these quadruplex at a concentration of 500 nM (data not shown).

The G4 ligand strategy Highly specific quadruplexes ligands have been designed and synthesized by Dr. Genevieve Prativiel (CNRS, Toulouse). These molecules are G4 specific structural probes. They weakly bind to the double-helix and single-stranded DNA but recognize very well all types of G4 with dissociation constants in the order of 100 nM.
In this second strategy, the G4 ligands are used as biotechnological tools for detecting the formation of G4s during the life cycle of HIV-1. As described above, we identified several HIV-1 sequences capable of adopting a G4 structure. The next step was therefore to test the effect of these ligands on the viral replication (Fig 5.8-D) and see if there is a correlation between the ability to stabilize the HIV-1 G4 *in vitro* (expressed as a ΔTm in $°C$) and the effect on viral infectivity *in cellulo* (Fig 5.8-E).

[4] forward sequences are generated for the sequences present in the reverse strand.

By testing a panel of 23 ligands, we have shown that there is a significant correlation between the affinity of a ligand for the G4 and its inhibitory potential (Fig 5.8-E). An increase of G4 stabilization by $7°C$ directly translates in an IC_{50} which is divided by 10. This correlation is true for a variety of compounds, independently of their chemical structure (Bisquinolinium, porphyrins and Salen).

Among this panel, the best compound has an IC_{50} of around 100 nM (under the same conditions, AZT ligand has an IC_{50} of 40 nM). For H2PYMA, AUT and TMPyP4 G4 ligands, FRET experiments performed in the presence of non G4 competitors demonstrated that they were much less specific: these compounds also bind to duplexes and single-strands. They turned out to be weak inhibitors of viral replication. We interpret this low efficiency by the "dilution" on non-specific targets. Cytotoxicity test on KB cells, A549, MCF7, MRC5, HCT116 and HeLa P4 showed no cytotoxic effect of these ligands on a period of 92 h and a concentration of up to 30 µM (Fig 5.8-F).

This correlation between stabilization *in vitro* and *in vivo* inhibition, associated to the high specificity of these ligands and the near absence of cytotoxicity on human cells, suggest that the observed inhibition is due to the recognition of quadruplex structures of the virus in the viral RNA or DNA. These compounds are thus able to detect the formation of G4 during the viral cycle. These ligands might become new antiviral molecules. We could not disclose the chemical formula of these ligands because of a patent being deposited on soon.

Conclusion We first analyzed the HIV sequence to find "natural" G4 DNA and RNA motifs present in its own genome and then demonstrated that these sequences have a strong antiviral activity. The observed inhibitory effects suggest that **these "viral" G4 act as decoys diverting viral or cellular proteins from their natural targets in the viral genome.** In a second step, we investigated whether small compounds could selectively bind to viral G4. The high specificity of these ligands and the absence of cytotoxicity suggests that **the inhibition observed at low concentrations is linked to specific antiviral effects of the 7 sequences we detected**.
Given these results, it is very likely that key steps of the viral cycle can be disrupted by our approaches, affecting the recognition of HIV's G4s by viral proteins or cellular factors. These potent inhibitory effects potentially pave the way for anti-HIV therapeutic applications constructed around these viral G4s.

5.2.3 Potential functions of these G4s

As stated above, it is interesting to note that G4-prone motifs detected by G4-Hunter in the HIV-1 genome are highly conserved despite the high genetic variability of the virus. This conservation suggest they play important functions in the HIV-1 replication cycle. Depending on the location of the G4 sequences, we can make some assumptions about their roles.

The C-ppt and GAG G-quadruplex sequences

The C-ppt sequence ($^{5'}$AAAGAAAAGGGGGGATT$^{3'}$) is located at the center of the genome over the integrase gene [192]. According to the literature, it is involved in the initiation of reverse transcription by forming the central initiation point of the (+) strand synthesis performed by the reverse transcriptase [193]. It also intervenes in the process of dimerization of the viral RNA as described previously. At the end of the synthesis of the second DNA strand (+) the C-ppt sequence (DNA) is found on a short overlapping strands called the "DNA flap" which is an essential element of the nuclear import step [194]. In a recent study, Bambara *et al* have suggested that recombination between the two RNA genomes and/or dimerisation could also be done via a bimolecular quadruplex using the C-ppt sequences of each of the viral RNAs [186]. They showed that the highly conserved G-rich sequence in the integrase gene near the central polypurine tract (C-ppt) dimerizes spontaneously at high ionic strength in the absence of protein. They found that the central regions of the viral RNA genomes (C-ppt) are likely to be maintained in proximity through dimer G-quartet formation during RNA template dimerization.

In another study, the same group also showed that a preferential recombination site involved the G4 GAG sequence [191]. This sequence was reported to form a G4 structure [192] and was able to fold into monomeric, dimeric and tetrameric G-quardruplexes depending on the cations employed [191].
These quadruplexes could facilitate the dimerisation and the exchange of material between the donor RNA and the recipient RNA. The formation of G4s could help RT in switching templates during synthesis of minus strand DNA, suggesting that the structure supports an increased recombination rate. Viral NCp7 accessory protein is known to facilitate the packaging of RNA, the reverse transcription and integration into the genome, but is also able to open intramolecular RNA structures to promote the bimolecular structures. With these chaperone qualities, NCp7 could promote the formation of a bimolecular G4 with a receiver RNA as was demonstrated by a Japanese group [195]. Significantly, the formation of an intermolecular G-quartet near the C-ppt and the gag gene would have a global effect on recombination. This would contribute to increased recombination rates in many other locations [186].

Nef gene G-quadruplex sequences

One attractive target for the anti-HIV therapy is the **Nef** protein. Nef is a small protein expressed early in the HIV-1 life cycle with at least two roles, (i) it is a fundamental factor for efficient viral replication and pathogenesis *in vivo*; (ii) it also facilitates virus replication and enhances viral infectivity *in vitro* [187]. Strikingly, viral isolates from some HIV-infected individuals that do not progress to AIDS exhibit either a deletion in the Nef gene or defective Nef alleles [196].

The 23-nt long conserved sequence that we detected has recently been studied by Perrone *et al* [187]. They showed that TMPyP4 G4 ligand changed the initial G4 topology and Tm of this sequence, indicating an effective interaction between the ligand and G4 Nef oligonucleotides. They treated HIV-1 infected cells with TMPyP4 and observed an efficient inhibition of Nef

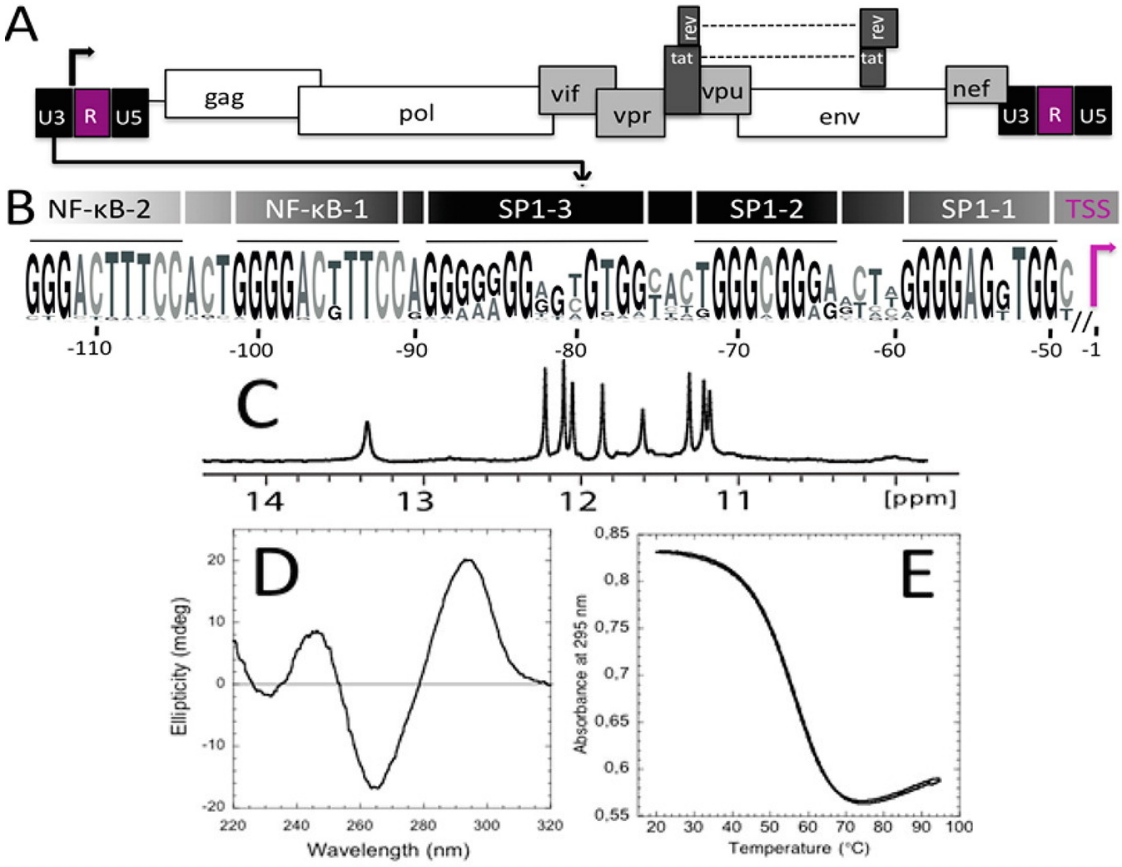


Figure 5.9: **Genomic structure of HIV-1 provirus and the conserved G-rich region of HIV-1 promoter.** (A) Genomic structure of HIV-1 provirus and (B) LOGO representation of the G-rich region of HIV-1 promoter generated by the weblogo software15 and based on an alignment of 1684 HIV-1 sequences from the HIV-1 database (www.hiv.lanl.gov). (C) Imino proton NMR spectrum recorded at 25°C at a concentration of 140 µM. (D) Circular dichroism spectrum recorded at 25°C at a concentration of 5 µM. (E) UV-melting profile recorded at 295 nm at a strand concentration of 140 µM. All experiments in this study were performed at 25°C in a buffer composed of 20 mM potassium phosphate pH 6.9 supplemented with 70 mM KCl. Adapted from [189].

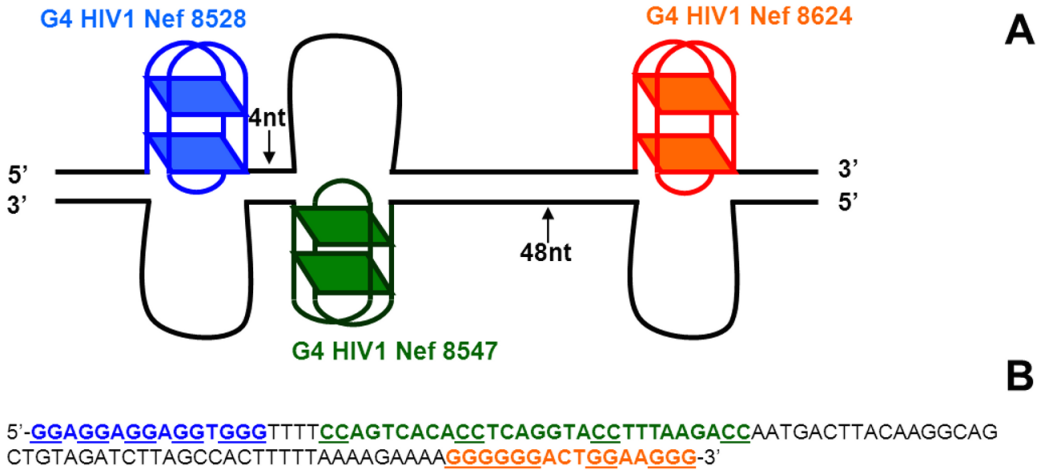


Figure 5.10: **Putative G-forming regions in the HIV-1 nef coding region.** (A) Scheme of G4 formation within the double-stranded DNA of the nef region: three G4 structures are shown in blue, green and red, respectively. The numbers of nts separating each G4 structure are indicated. (B) Nucleotide sequence of the nef coding region where three putative G4 sequences were identified. Two sequences (blue and red) were identified on the non-coding strand, thus the reverse complementary sequence is shown on the upper strand (in green). Adapted from [187].

protein expression. Moreover, G4 stabilization within the Nef coding region of the viral genome may reduce Nef expression through inhibition of transcription that directly generates mRNAs, as well as overall transcription for production of new copies of the RNA genome. Perrone *et al* results may open a new avenue in the development of antiviral compounds with an unprecedented mechanism of action.

The sequence located on the Nef gene at the boundary of the U3 region contains the 3'PPT sequence ($^{5'}$AAAGAAAAGGGGGGACT$^{3'}$). It is involved in the initiation of reverse transcription by forming the first point of (+) strand synthesis performed by the reverse transcriptase. The 93del aptamer was originally discovered by a SELEX approach against the RNase H domain of the reverse transcriptase, demonstrating that this viral enzyme binds strongly to G4. In addition, the C-ppt sequences and PPT, which form stable G4, are not degraded by RNase H. This persistence is potentially linked to the recognition of the C-ppt and PPT G4s by this enzyme, which is essential for starting the synthesis of the DNA strand (+).

G-quadruplexes in the U3 region of HIV-1 promoter

During reverse transcription, the single-stranded viral RNA is converted into double-stranded DNA called provirus. In order to regenerate the regulatory sequences required for integration and expression of the provirus, an intramolecular recombination happens during reverse transcription between the 5' U5R and 3' U3R extremities of the RNA genome [197]. The 9200 nucleotides HIV RNA genome is converted into a double-stranded DNA provirus consisting of 9700 bp and bordered by the two LTRs regulators (Fig 5.9-A). The G4s detected in the U3 region of the RNA are located in both the 5' LTR and 3'LTR of the DNA provirus.

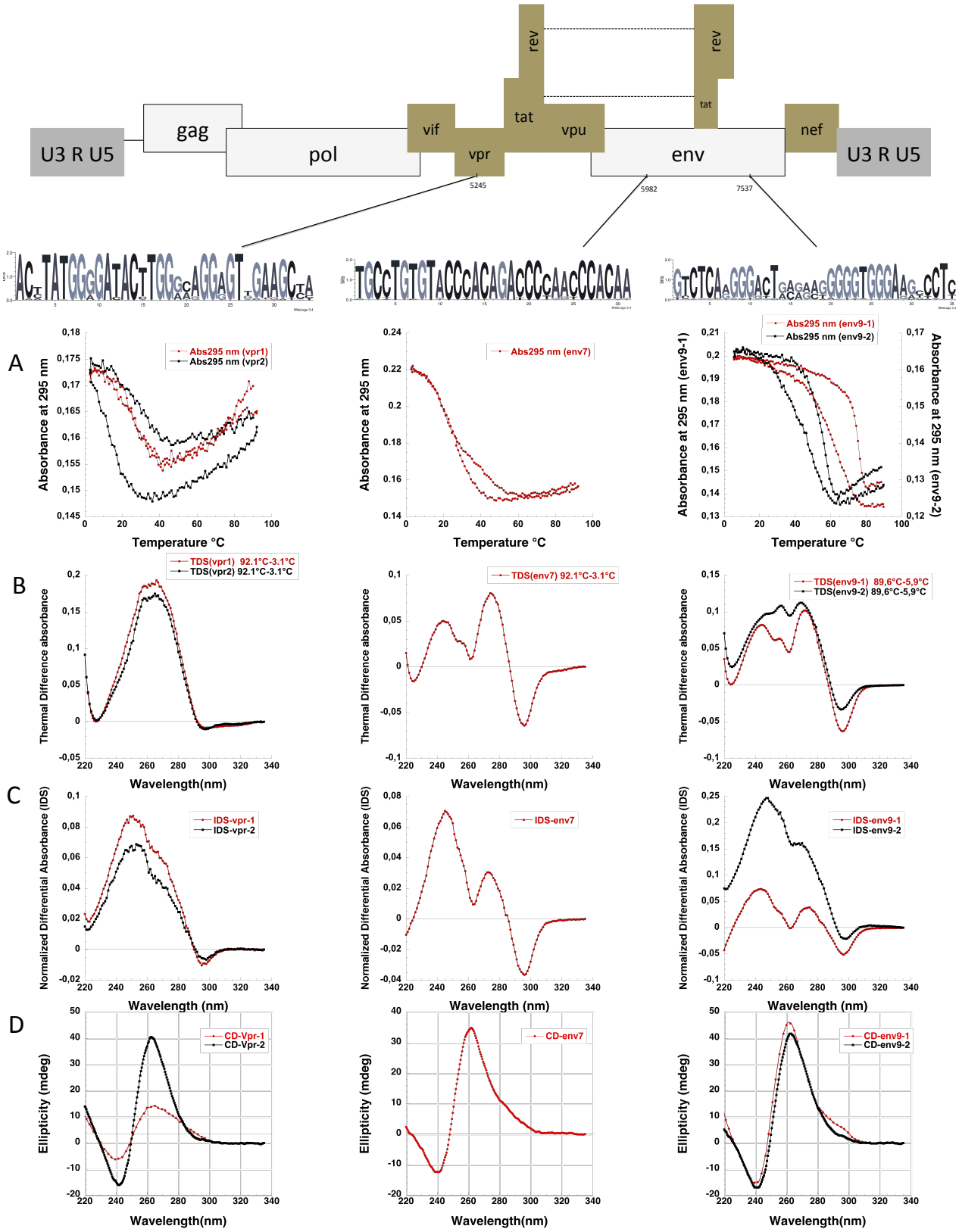


Figure 5.11: **New Potential quadruplex sequences in the HIV-1 genome.** Distribution of vpr (1 & 2), env7, env9 (1 & 2) PQS in the HIV-1 genome. (top) HIV-1 genome and LOGO representation. (A) UV-melting profile recorded at 295 nm at a strand concentration of 4 µM. (B-C-D) TDS, IDS, CD assays. All experiments in this study were performed in a buffer composed of 100 mM KCl. (C-D) IDS and CD were performed at 25°C.

The 5' LTR serves as a promoter for the entire viral genome, while the 3' LTR has the stop codon and provides the polyadenylation site for the new viral RNA transcript. At the 5' LTR, it overlaps with the binding sites of SP1 and NFκB transcription factors, only a few nucleotides upstream of the transcription start site and the TATA box and regulates the promoter activity. This promoter region contains a G-rich sequence of 50 nucleotides upstream from the transcription-starting site (TSS). Situated from -57 to -78 bases from the TSS, *in vitro* assays showed their ability to form an antiparallel stable G4 structure with a Tm of $60°C$ (Fig 5.9-D-E). The imino proton NMR of the HIV-PRO1 sequence exhibits eight resolved imino peaks in the 10-12 ppm region (Fig 5.9-C). This point out to a two G-tetrad structure.

Recently, Perrone *et al* suggested that the formation of a G4 structure in this region is able to repress the activity of the HIV-1 promoter [198]. Different studies [198, 189] showed that G4 ligands are able to inhibit HIV-1 infectivity, which suggest that G4s can be targeted to treat AIDS. We also demonstrated that specific G4 ligands are able to strongly inhibit the virus. The mechanisms of action of these new inhibitors are being analyzed, but we can hypothesize that at least part of this inhibition might be related to HIV-1 promoter repression. These findings reveal the possibility of inhibiting the HIV-1 LTR promoter by G-quadruplex-interacting small molecules, providing a new pathway to development of anti-HIV-1 drugs with an unprecedented mechanism of action.

5.2.4 New conserved G4 sequences in *vpr* and *env* regions

G-quadruplexes in the *env* gene

Human immunodeficiency virus type 1 is completely dependent upon the Env protein to enter cells. The HIV-1 env gene encodes the only surface-expressed viral protein Env. Env is a glycoprotein of 160 kD (gp160), which is exclusively required for binding and entry into host cells. After translation, gp160 is cleaved by cellular proteases into mature proteins gp120 and gp41, that are non-covalently linked to form a single subunit of a trimeric "spike" on the virion surface. The C-terminal subunit, gp41, contains three domains, (i) a cytoplasmic domain (inside the viral membrane), (ii) a membrane-spanning domain and (iii) an extracellular domain, which mediates the conformational change needed for fusion. The N-terminal subunit, gp120, is completely outside the viral membrane and organized into five conserved regions (C1-C5) interspersed with five variable regions (V1-V5). The Env protein is an entry machine, built to bind to CD4 receptors of the host cell, undergo a series of conformational changes, fuse the cell and viral membranes and deliver the viral core to the cytoplasm of the cell. [199]. In regard to transmission, the viral envelop protein is not only responsible for viral entry but also modulates certain functions of host cells that facilitate infection. Mutations in the viral envelop proteins may affect viral infectivity through different mechanisms. Mutations in the CD4 binding site (CD4bs) of gp120 may cause the virus to become non-infectious [200, 201].

Using G4-Hunter we identified two potential quadruplex sequences in this gene. One sequence on the (+) strand and the second on the (-) strand respectively. It is the first time that these two sequences are reported as conserved PQS G4 structure in the env gene within the HIV-1 genome.

According to the two logos representation we generated three sequences that appear to be the most frequent (Table 5.2). One conserved sequences named env7 in the position 5982 nt and two sequences named env9-1 and env9-2, which represent the most occurred PQS in the region 7537 nt. the biophysical assays confirm the folding of these sequences to a parallel G4 structures (Fig 5.11).

The importance of these sequences within a viral context is assessed by the bases conservation and the formation of the G4 structures in this genomic region could affect polymerase processing and hence influence replication and transcription of the env gene.

G4 in the *vpr* gene

In addition to the retroviral Gag, Pol and Env proteins, HIV-1 produces the Vpr accessory protein with an average length of 96 amino acids and molecular weight of 12.7 kD. The accessory proteins are dispensable for viral proliferation in many *in vitro* systems but often necessary for viral replication and pathogenesis *in vivo* [202]. Vpr is highly conserved among HIV, simian immunodeficiency viruses (SIV) and other lentiviruses [203, 204]. Additionally, it is required for full pathogenesis *in vivo* by a mechanism that is poorly understood. It displays several distinct activities in host cells. These include cytoplasmic-nuclear shuttling [205], induction of cell cycle G2 arrest [206, 207] and cell killing [208].

These three Vpr-specific activities have been demonstrated in a wide variety of eukaryotic cells ranging from yeast to humans, indicating that Vpr most likely affects highly conserved cellular processes [209].

G4-hunter identified a conserved G-rich sequence around the position 5245 nts of the vpr gene, where the majority of Gs bases were highly conserved. Two main frequent sequences were selected (vpr-1 and vpr-2) and for the *in vitro* tests both of them do fold into G4 structures with a parallel topology according to the CD spectrum (Fig 5.11). The vpr activities are mediated by interactions with different partners and the changes in the accessibility of vpr regulates these interactions. Thereby we suppose that the formation of G4 sequences allows different conformations, which can alter these interaction, the vpr functions and the the viral infectivity.

5.3 Hunting new G-quadruplexes in Ebola and Marburg viruses

The Outbreaks of haemorrhagic fever caused by Ebola virus in West Africa, which began December 2014 in Guinea and has since spread to Liberia and Sierra Leone, has already caused thousands cases. Ebola virus and Marburg virus (EBOV and MBGV) are the only members of the Filoviridae family. Both originated from Africa and are a causative agent of viral haemorrhagic fever, muscle pain followed by vomiting, diarrhoea, kidney and liver damage.

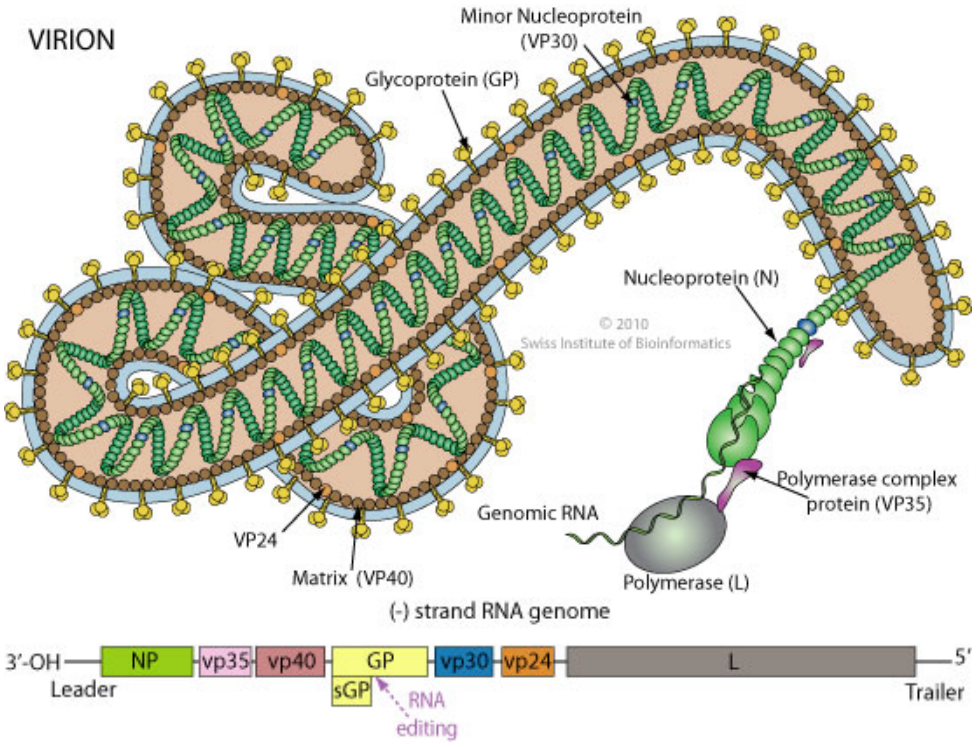


Figure 5.12: **Presentation of filamentous 970 nm-long Ebolavirus** (Diameter is about 80 nm). Ebola has a negative-stranded RNA linear genome, about 18-19 kb in size, which encodes for seven proteins. Adapted from Viral Bioinformatics Resource Center (VBRC).

Ebola and Marburg are very similar in terms of morphology, genome organization and protein composition (Fig. 5.12). They are filamentous, negative-sense, nonsegmented single-stranded RNA viruses which cause frequently lethal hemorrhagic fevers in humans and nonhuman primates [210]. Morphologically, the virus consists of a linear genome entirely enclosed in an envelop, which is coated by the membrane glycoprotein, organized in homotrimeric. All of the filovirus genomes sequenced to date are approximately 19 Kb negative-sense, single-stranded RNA and all encodes seven structural proteins and one non-structural protein (soluble glycoprotein) (Fig. 5.12). The seven functional proteins encoded in EBOV are designated as NP (nucleoprotein), viral proteins (VP) VP24 (membrane-associated protein), VP30, VP35 (both polymerase matrix proteins), VP40 (matrix protein), L (RNA polymerase) and GP (glycoprotein) [211, 212].

Table 5.4: *Filoviridae* family, all the species from Ebolavirus and Marburgvirus genus.

Family	Genus	Species	Isolate	accession	Bases
Filoviridae	*Ebolavirus*	*Cote d'Ivoire*	Cote d'Ivoire ebolavirus strain Cote d'Ivoire	FJ217162	18935
		Reston ebolavirus	Reston ebolavirus strain Pennsylvania Lyon	AY769362	18895
			Reston ebolavirus strain Pennsylvania-Groseth	NC_004161	18891
			Reston ebolavirus strain Reston	AB050936	18890
		Sudan ebolavirus	Sudan ebolavirus strain Gulu	NC_006432	18875
			Sudan ebolavirus strain Yambio	EU338380	18875
		Unclassified ebolavirus	Bundibugyo ebolavirus strain Uganda	FJ217161	18940
		Zaire ebolavirus	Zaire ebolavirus strain Mayinga Russia	EU224440	18959
			Zaire ebolavirus strain Mayinga Wilson	AF499101	18960
			Zaire ebolavirus strain Mayinga-Volchkov	NC_002549	18959
			Zaire ebolavirus strain Mayinga-Zaire	AY142960	18959
			Zaire ebolavirus strain Mayinga-Zaire8mc	AF272001	18959
			Zaire ebolavirus strain Zaire 1995	AY354458	18961
	Marburgvirus	*Lake Victoria marburgvirus*	Lake Victoria marburgvirus - Ci67 strain Germany_ Marburg	EF446132	19112
			Lake Victoria marburgvirus - Ravn strain Kenya	EF446131	19114
			Lake Victoria marburgvirus Angola2005 strain Ang0126	DQ447656	19114
			Lake Victoria marburgvirus Angola2005 strain Ang0214	DQ447657	19114
			Lake Victoria marburgvirus Angola2005 strain Ang0215	DQ447658	19114
			Lake Victoria marburgvirus Angola2005 strain Ang0754	DQ447659	19114
			Lake Victoria marburgvirus Angola2005 strain Ang0998	DQ447660	19114
			Lake Victoria marburgvirus Angola2005 strain Ang1379c	DQ447653	19114
			Lake Victoria marburgvirus Angola2005 strain Ang1381	DQ447654	19114
			Lake Victoria marburgvirus Angola2005 strain Ang1386	DQ447655	19114
			Lake Victoria marburgvirus DRC1999 strain 05DRC99	DQ447651	19114
			Lake Victoria marburgvirus DRC1999 strain 07DRC99	DQ447650	19114
			Lake Victoria marburgvirus DRC1999 strain 09DRC99	DQ447652	19114
			Lake Victoria marburgvirus strain Musoke	DQ217792	19111
			Lake Victoria marburgvirus strain Popp	Z29337	19112
			Lake Victoria marburgvirus strain pp3 guinea pig lethal variant	AY430365	19113
			Lake Victoria marburgvirus strain pp4 guinea pig nonlethal variant	AY430366	19112
			Lake Victoria marburgvirus strain Ravn	DQ447649	19114
			Marburg virus strain M/S.Africa/Johannesburg/1975/Ozolin	AY358025	19151
			Marburg virus strain Musoke	Z12132	19104

Genetic and antigenic characterization of Ebola virus (EBOV) isolates during human outbreaks has led to the identification of four subtypes: E. Sudan, E. Zaire, E. Ivory Coast and E. Reston [213] (Table 5.4). In contrast to Ebola Reston, which originates in Asia and has never been reported to cause human disease [214], the other three subtypes circulate on the African subcontinent and are pathogenic for humans, causing a specific febrile hemorrhagic disease. After an incubation period of about a week, victims rapidly develop high fever, diarrhea, vomiting, respiratory disorders, hemorrhaging and impaired immunity. Death ensues within few days. Fatal infections are also characterized by progressively increasing systemic viral titers and cytokines, consistent with a model in which host innate and adaptive immune responses are unable to control infection, while the inflammatory response becomes overactivated, causing disease [215].

The fatality rates are about 80% with E. Zaire [216] while fatalities in epidemics caused by the Sudan species have been in the range of 50–60% [217]. The broad-spectrum anti-viral agent ribavirin, a trizole nucleoside affective against multiple pathogenic RNA viruses, is not active against filoviruses [218]. Other small molecules have accorded a high-degree of protection against filoviruses [219]. Recentlly, the work of Warren K.W. *et al* show that a novel synthetic adenosine analogue BCX4430, inhibits infection of distinct filoviruses in human cells. This by acting as a non-obligate RNA chain terminator. BCX4430 was well tolerated, producing no overt signs of toxicity [220].

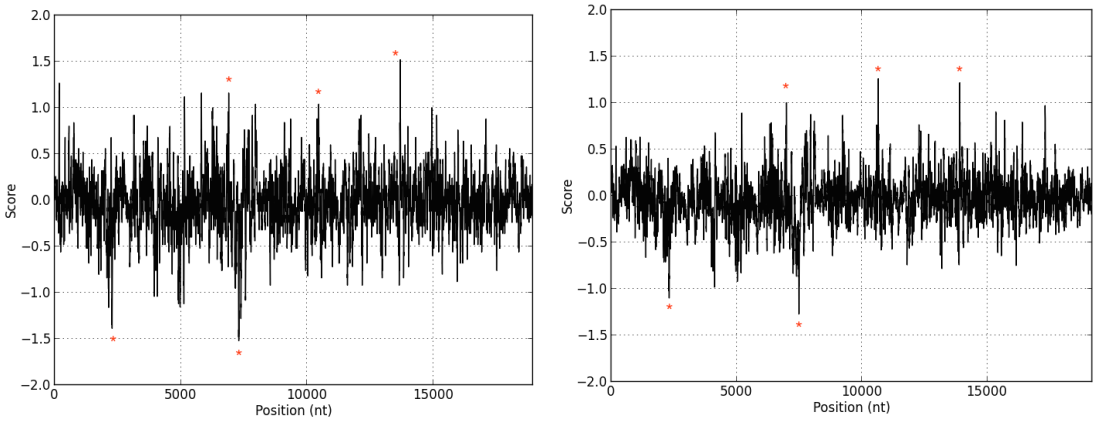


Figure 5.13: **G4Score distribution in aligned Ebola genomes.** Left, alignment of 99 *Ebola Zaire* sequences, right alignment of 13 Ebola genome. Stars represent the five conserved sequences in the 112 genomes.

Objectives

Like HIV, we think that G-quadruplexes might interfere with other viral, of which Ebola and Marburg, replication cycles. Using our G4-hunter algorithm, we searched for a conserved G4 sequences in 20 aligned genome of Marburg and 13 aligned Ebola genome using the parameters: window of 25 nt and threshold of 1. These genomes were extracted from Viral Bioinformatics Resource Center (VBRC)[5]. The work of S.K. Gire [221] provide a liste of 99 new Ebola genomes from 78 patients in Sierra Leone. This work allows us to enrich our Ebola dataset to 112 sequences (Fig. 5.13). The searching of G4s in the 99 Ebola genomes identify more then 16 conserved PQS among the Ebola Zaire Species (25 nt windows and threshold of 1). Mixing this genomes to previous Ebola genomes extracted from VBRC only 5 sequences are conserved.

Where as Marburg analysis identified 17 conserved PQS. The *in vitro* validation demonstrated the folding of 4 Ebola sequences into stable G-quadruplexes. In the Marburg case, 10 of the 17 PQS fold into stable G4 (Annexe II Table 28).

We demonstrated the presence of G4 sequences in the genomes of these two viruses which, as in the case of HIV, are potentially involved in the regulation of their replication cycles. In future projects, we plan to test the decoy and the G4 ligands strategies on each of the viruses, as we have powerful G4 ligands and 15 new G-quadruplexes.

[5]Not more availeble.

5.4 Patent 1: Nucleic acids acting as decoys for the treatment of lentivirus infection.

S.Amrane[1,2], ML. Andreola[3], A. Bedrat[1,2] & JL Mergny[1,2]
EP14305763.6, deposit Mai 23^{th} 2014.

1. Université de Bordeaux, ARNA Laboratory, F-33000 Bordeaux, France;
2. Inserm U869, IECB, F-33600 Pessac, France;
3. Université de Bordeaux UMR-CNRS 5234, F-33076 Bordeaux, France.

The present invention relates to nucleic acid sequences which are fragments of HIV-1 genome. The nucleic acids of the present invention are use as decoys for the treatment of lentivirus infection. In particular, the present invention relates to nucleic acids capable of forming at least one G-quadruplex domain and that are capable of inhibiting the replication of at least one lentivirus, such as HIV-1.

NUCLEIC ACIDS ACTING AS DECOYS FOR THE TREATMENT OF LENTIVIRUS INFECTION

FIELD OF THE INVENTION:

The present invention relates to nucleic acid sequences which are fragments of HIV-1 genome. The nucleic acids of the present invention are use as decoys for the treatment of lentivirus infection.

BACKGROUND OF THE INVENTION:

Human immunodeficiency virus (HIV-1) is a retrovirus responsible of a global pandemic inducing a deficiency of the immune system causing AIDS. HIV-1 retrovirus infects cells that carry CD4 and one of the chemokine receptors CCR5 or CXCR4[1]. After infection, the two HIV-1 single-stranded RNAs are reverse transcribed by the viral reverse transcriptase into double-stranded DNA. The viral DNA is then integrated into the genome of the infected cell. The host cell machinery transcribes the viral genes, new viral proteins are synthesized, and new viruses are finally assembled. At the end of the 1990, the setting of an antiviral therapy targeting different enzymes of the viral cycle was a tremendous step forward in the battle against AIDS. However this treatment did not succeed into the definitive eradication of the virus and due to some mutations in the genome of the virus, resistance against these molecules can occur. Indeed, according to UNAIDS, 34 million people are currently infected by HIV-1. The discovery of new anti-viral strategies is still an important issue.

DNA or RNA sequences containing guanine tracts are able to adopt non-canonical four-stranded structures called G-quadruplexes (G4s)[2]. The core of the G4 is based on the stacking of 2 or more G-tetrads. Each tetrad is a planar association of four guanines held together by eight hydrogen bonds and coordinated with a central Na^+ or K^+ cation. Unlike the canonical duplex, G4s form a very polymorphic family of globularly shaped nucleic acid structures. G4s can be thermally stable with melting temperatures typically above 40°C under near physiological conditions,. Genome scale bioinformatics analysis showed a significant enrichment of these sequences in regulatory elements of the human genome such as telomeres and oncogenes. In vivo studies, using specific G4 probes[3,4] strongly suggests the formation of G4s in cells. The implication of G-quadruplexes in virology[5] is also the subject of recent investigations in the papilloma, Epstein-Barr and SARS viruses. However, use of G-

quadruplexes derived from the own virus genome sequence for the treatment of HIV-1 infection has never been suggested in the prior art

SUMMARY OF THE INVENTION:

The present invention relates to nucleic acid sequences which are fragments of HIV-1 genome. The nucleic acids of the present invention are use as decoys for the treatment of lentivirus infection. In particular, the present invention relates to nucleic acids capable of forming at least one G-quadruplex domain and that are capable of inhibiting the replication of at least one lentivirus, such as HIV-1.

DETAILED DESCRIPTION OF THE INVENTION:

Several studies show that the viral proteins are able to recognize G4 structures with high affinity and specificity[6,7]. This striking trend prompted the inventors to search for G4 forming sequences in the HIV-1 genome that could be recognised by the viral proteins *in vivo*. Using a bioinformatics approach they identified three very conserved G4 forming sequences that are involved in key steps of the HIV-1 replication cycle. The inventors purchased and tested these oligonucleotides in a viral infectivity test and they showed that these sequences are very potent HIV-1 inhibitors with effects in the nanomolar range. These G4s might therefore act as decoys that trap crucial proteins involved in the recognition of the same sequences present in the HIV-1 genome. Accordingly, the present invention relates to nucleic acids capable of forming at least one G-quadruplex domain and that are capable of inhibiting the replication of at least one lentivirus, such as HIV-1.

The term "lentivirus" as used herein, refers to human immunodeficiency virus-1 (HIV-1); human immunodeficiency virus-2 (HIV-2); simian immunodeficiency virus (SIV); feline immunodeficiency virus (FIV) and equine immunodeficiency virus (EIV)

As used herein the term "G-quadruplex domain" refers to any guanosine-rich oligonucleotide sequence capable of forming G-tetrads, each of which is a square arrangement of guanines stabilized by Hoogsteen hydrogen bonding, and which may be further stabilized by the presence of a monovalent cation (especially potassium) in the center of the tetrads, without further limitation as to sequence. A G-quadruplex structure may include 2, 3, 4, 5 or more tetrads. A G-quadruplex structure may be formed of DNA, RNA or a modified nucleic acid such as an LNA or a PNA. Resources including algorithms for identifying and predicting

sequences which have the capacity to form G-quadruplexes are readily available, for example online and in QUADRUPLEX NUCLEIC ACIDS, Neidle & Balasubramanian (Eds.) 2006.

In all subsequent paragraphs, all nucleic acid sequences are listed in the 5' to 3' direction.

In some embodiments, the present invention relates to a nucleic acid (NA3) having the general formula (I) of :

$L3_1$-$(G)_{4-6}$-$L3_2$-$(G)_{4-6}$-$L3_3$-$(G)_{3-4}$-$L3_4$ (I)

wherein

- $(G)_n$ represents a sequence of n guanosines
- $L3_1$ represents a sequence of 0 to 4 nucleotides
- $L3_2$ represents a sequence of 2 to 5 nucleotides
- $L3_3$ represents a sequence of 5 to 10 nucleotides
- $L3_4$ represents a sequence of 0 to 4 nucleotides

As used herein the terms "nucleotide" has its general meaning in the art and includes, but is not limited to, a natural nucleotide, a synthetic nucleotide, or a nucleotide analogue. The nucleoside phosphate may be a nucleoside monophosphate, a nucleoside diphosphate or a nucleoside triphosphate. The sugar moiety in the nucleoside phosphate may be a pentose sugar, such as ribose, and the phosphate esterification site may correspond to the hydroxyl group attached to the C-5 position of the pentose sugar of the nucleoside. A nucleotide may be, but is not limited to, a deoxyribonucleoside triphosphate (dNTP) or a ribonucleoside triphosphate (NTP). The nucleotides may be represented using alphabetical letters (letter designation), as described in Table A. For example, A denotes adenosine (i.e., a nucleotide containing the nucleobase, adenine), C denotes cytosine, G denotes guanosine, and T denotes thymidine. W denotes either A or T/U, and S denotes either G or C. N represents a random nucleotide (i.e., N may be any of A, C, G, or T/U). As used herein, the term "nucleotide analogue" refers to modified compounds that are structurally similar to naturally occurring nucleotides. The nucleotide analogue may have an altered phosphorothioate backbone, sugar moiety, nucleobase, or combinations thereof. Generally, nucleotide analogues with altered nucleobases confer, among other things, different base pairing and base stacking properties.

Nucleotide analogues having altered phosphate-sugar backbone (e.g., PNA, LNA, etc.) often modify, among other things, the chain properties such as secondary structure formation. At times in the instant application, the terms "nucleotide analogue," "nucleotide analogue base," "modified nucleotide base," or "modified base" may be used interchangeably.

The term "nucleic acid" refers to a macromolecule composed of chains of monomeric nucleotides, which forms a structure that demonstrates a biological function and may also carry genetic information. As is the case with polynucleotides, nucleic acids encompass DNA and RNA in double- and single-stranded forms, and also encompass nucleic acids wherein the bases are a modified form of either type of nucleotide, including LNAs, PNAs, HNAs, GNAs and TNAs. It will be understood that nucleic acids that comprise a nucleotide sequence as disclosed herein also encompass those nucleic acids wherein thymidine (T) may be replaced in the sequence by uracil (U), such as when uracil (U) in an RNA sequence replaces thymidine (T) in a corresponding DNA sequence.

As used herein, the term "oligonucleotide" means any macromolecule that is a polymer of monomeric nucleotides. The nucleotiode bases may be either ribonucleotides or deoxynucleotides or a modified form of either type of nucleotide, including those that display increased thermal stabilities when hybridized to complementary DNAs or RNAs as compared to unmodified DNA:DNA and DNA:RNA pairs. Such modified nucleotides include morpholino and locked nucleic acids (LNAs), peptide nucleic acids (PNAs), 1', 5'-anhydrohexitol nucleic acids (HNAs), glycol nucleic acids (GNAs) and threose nucleic acids (TNAs), all of which are characterized by changes to the backbone of the molecule and are capable of folding to form quadruplex structures. The term "polynucleotide" also is meant to encompass single and double stranded forms of nucleotides. Polynucleotides that comprise a nucleotide sequence as disclosed herein also encompass those polynucleotides wherein thymidine (T) may be replaced in the sequence by uracil (U), such as when uracil (U) in an RNA sequence replaces thymidine (T) in a corresponding DNA sequence, inasmuch as one of the four major bases in RNA is uracil (U) rather than thymidine (T) as in DNA.

In some embodiments, $L3_1$ represents AA.

In some embodiments, $L3_2$ represents ATT or AUU.

In some embodiments, $L3_3$ represents TACAGTGCA or UACAGUGCA.

In some embodiments, $L3_4$ represents AA.

In some embodiments, the nucleic acid (NA3) is represented by SEQ ID NO: 11 or SEQ ID NO:12.

In some embodiments, the present invention relates to a nucleic acid (NA9) having the general formula (II) of:

$L9_1$-$(G)_{4-6}$-$L9_2$-$(G)_{2-3}$-$L9_3$-$(G)_{2-4}$-$L9_4$ (II)

Wherein

- $(G)_n$ represents a sequence of n guanosines
- $L9_1$ represents a sequence of 0 to 4 nucleotides
- $L9_2$ represents a sequence of 2 to 4 nucleotides
- $L9_3$ represents a sequence of 1 to 4 nucleotides
- $L9_4$ represents a sequence of 0 to 4

In some embodiments, $L9_1$ represents AA.
In some embodiments, $L9_2$ represents ACT or ACU.
In some embodiments, $L9_3$ represents AT or AU.
In some embodiments, $L9_4$ represents AA.

In some embodiments, the nucleic acid (NA9) is represented by SEQ ID NO:10 or SEQ ID NO:13.

In some embodiments, the present invention relates to a nucleic acid (NA10) having the general formula (III) of:

$L10_0$-$(G)_{3-4}$-$L10_1$-$(G)_{3-4}$-$L10_2$-$(G)_{2-3}$-$L10_3$-$(G)_{2-3}$-$L10_4$-$(G)_{3-4}$-$L10_5$-$(G)_{3-4}$-$L10_6$-$(G)_{3-4}$-$L10_7$- $(G)_{2-3}$-$L10_8$ (III)

Wherein

- $(G)_n$ represents a sequence of n guanosines
- $L10_0$ represents a sequence of 0 to 4 nucleotides
- $L10_1$ represents a sequence of 7 to 10 nucleotides

- $L10_2$ represents a sequence of 1 to 3 nucleotides
- $L10_3$ represents a sequence of 1 to 4 nucleotides
- $L10_4$ represents a sequence of 2 to 4 nucleotides
- $L10_5$ represents a single nucleotide
- $L10_6$ represents a sequence of 2 to 5 nucleotides
- $L10_7$ represents a sequence of 2 to 5 nucleotides
- $L10_8$ represents a sequence of 0 to 4 nucleotides

In some embodiments, $L10_1$ represents ACTTTCC or ACUUUCC.
In some embodiments, $L10_2$ represents A.
In some embodiments, $L10_3$ represents CGT or CGU.
In some embodiments, $L10_4$ represents CCT or CCU.
In some embodiments, $L10_5$ represents C.
In some embodiments, $L10_6$ represents ACT or ACU.
In some embodiments, $L10_7$ represents AGT or AGU.

In some embodiments, the nucleic acid (NA10) is represented by SEQ ID NO:1 (GGGACTTTCCGGGGAGGCGTGGCCTGGGCGGGACTGGGGAGTGGC).

In some embodiments, the present invention relates to a nucleic acid (NA10.1) having the general formula (IV) of:

$$L10_0\text{-}(G)_{3\text{-}4}\text{-}L10_1\text{-}(G)_{3\text{-}4}\text{-}L10_2\text{-}(G)_{2\text{-}3}\text{-}L10_3\text{-}(G)_{2\text{-}3}\text{-}L10_4 \quad (IV)$$

Wherein

- $(G)_n$ represents a sequence of n guanosines
- $L10_0$ represents a sequence of 0 to 4 nucleotides
- $L10_1$ represents a sequence of 7 to 10 nucleotides
- $L10_2$ represents a sequence of 1 to 3 nucleotides
- $L10_3$ represents a sequence of 1 to 4 nucleotides
- $L10_4$ represents a sequence of 0 to 4 nucleotides

In some embodiments, $L10_1$ represents ACTTTCC or ACUUUCC.
In some embodiments, $L10_2$ represents A.

In some embodiments, $L10_3$ represents CGT or CGU.

In some embodiments, $L10_4$ represents C.

In some embodiments, the nucleic acid (NA10.1) is represented by SEQ ID NO:2 or SEQ ID NO:3.

In some embodiments, the present invention relates to a nucleic acid (NA10.2) having the general formula (V) of:

$$L10_1\text{-}(G)_{3\text{-}4}\text{-}L10_2\text{-}(G)_{2\text{-}3}\text{-}L10_3\text{-}(G)_{2\text{-}3}\text{-}L10_4\text{-}(G)_{3\text{-}4}\text{-}L10_5\text{-}(G)_{3\text{-}4}\text{-}L10_6 \quad (V)$$

Wherein

- $L10_1$ represents a sequence of 0 to 4 nucleotides
- $L10_2$ represents a sequence of 1 to 3 nucleotides
- $L10_3$ represents a sequence of 1 to 4 nucleotides
- $L10_4$ represents a sequence of 2 to 4 nucleotides
- $L10_5$ represents a single nucleotide
- $L10_6$ represents a sequence of 0 to 4 nucleotides

In some embodiments, $L10_1$ represents A.

In some embodiments, $L10_2$ represents A.

In some embodiments, $L10_3$ represents CGT or CGU.

In some embodiments, $L10_4$ represents CCT or CCU.

In some embodiments, $L10_5$ represents C.

In some embodiments, the nucleic acid (NA10.2) is represented by SEQ ID NO:4 or SEQ ID NO:5.

In some embodiments, the present invention relates to a nucleic acid (NA10.3) having the general formula (VI) of:

$$L10_3\text{-}(G)_{2\text{-}3}\text{-}L10_4\text{-}(G)_{3\text{-}4}\text{-}L10_5\text{-}(G)_{3\text{-}4}\text{-}L10_6\text{-}(G)_{3\text{-}4}\text{-}L10_7 \quad (VI)$$

Wherein

- $(G)_n$ represents sequence of n guanosines
- $L10_3$ represents a sequence of 0 to 4 nucleotides
- $L10_4$ represents a sequence of 2 to 4 nucleotides
- $L10_5$ represents a single nucleotide
- $L10_6$ represents a sequence of 2 to 5 nucleotides
- $L10_7$ represents a sequence of 0 to 4 nucleotides

In some embodiments, $L10_3$ represents T or U.
In some embodiments, $L10_4$ represents CCT or CCU.
In some embodiments, $L10_5$ represents C.
In some embodiments, $L10_6$ represents ACT or ACU.

In some embodiments, the nucleic acid (N10.3) is represented by SEQ ID NO:6 or SEQ ID NO:7.

In some embodiments, the present invention relates to a nucleic acid (NA10.4) having the general formula (VII) of:

$$L10_4\text{-}(G)_{3\text{-}4}\text{-}L10_5\text{-}(G)_{3\text{-}4}\text{-}L10_6\text{-}(G)_{3\text{-}4}\text{-}L10_7\text{-}(G)_{2\text{-}3}\text{-}L10_8 \quad \text{(VII)}$$

Wherein
- $(G)_n$ represents a sequence of n guanosines
- $L10_4$ represents a sequence of 0 to 4 nucleotides
- $L10_5$ represents a single nucleotide
- $L10_6$ represents a sequence of 2 to 5 nucleotides
- $L10_7$ represents a sequence of 2 to 5 nucleotides
- $L10_8$ represents a sequence of 0 to 4 nucleotides

In some embodiments, $L10_4$ represents CCT or CCU.
In some embodiments, $L10_5$ represents C.
In some embodiments, $L10_6$ represents ACT or ACU.
In some embodiments, $L10_7$ represents AGT or AGU.

In some embodiments, the nucleic acid (NA10.4) is represented by SEQ ID NO:8 or SEQ ID NO:9.

For use in the instant invention, the nucleic acids of the present invention can be synthesized de novo using any of a number of procedures well known in the art. Chemical synthesis can be performed by a variety of automated nucleic acid synthesizers available in the market. These nucleic acids may be referred to as synthetic nucleic acids. Alternatively, the nucleic acids of the present invention can be produced on a large scale in plasmids. The nucleic acids of the present invention can be prepared from existing nucleic acid sequences using known techniques, such as those employing restriction enzymes, exonucleases or endonucleases.

A further object of the present invention relates to a method of treating a lentivirus infection in a subject in need thereof comprising administering the subject with a therapeutically effective amount of at least one nucleic acid of the present invention.

As used herein, the term “subject” denotes a mammal, such as a rodent, a feline, a canine, a equine and a primate. Preferably, a subject according to the invention is a human.

In some embodiments, the method of the present invention is particularly suitable for the treatment of HIV-1 infections.

By a "therapeutically effective amount" is meant a sufficient amount of nucleic acid of the present invention to treat and/or to prevent lentivirus infections (e.g. HIV-1 infections) at a reasonable benefit/risk ratio applicable to any medical treatment. It will be understood that the total daily usage of the compounds and compositions of the present invention will be decided by the attending physician within the scope of sound medical judgment. The specific therapeutically effective dose level for any particular subject will depend upon a variety of factors including the disorder being treated and the severity of the disorder; activity of the specific compound employed; the specific composition employed, the age, body weight, general health, sex and diet of the subject; the time of administration, route of administration, and rate of excretion of the specific compound employed; the duration of the treatment; drugs used in combination or coincidental with the specific polypeptide employed; and like factors well known in the medical arts. For example, it is well within the skill of the art to start doses

of the compound at levels lower than those required to achieve the desired therapeutic effect and to gradually increase the dosage until the desired effect is achieved. However, the daily dosage of the products may be varied over a wide range from 0.01 to 1,000 mg per adult per day. Preferably, the compositions contain 0.01, 0.05, 0.1, 0.5, 1.0, 2.5, 5.0, 10.0, 15.0, 25.0, 50.0, 100, 250 and 500 mg of the active ingredient for the symptomatic adjustment of the dosage to the subject to be treated. A medicament typically contains from about 0.01 mg to about 500 mg of the active ingredient, preferably from 1 mg to about 100 mg of the active ingredient. An effective amount of the drug is ordinarily supplied at a dosage level from 0.0002 mg/kg to about 20 mg/kg of body weight per day, especially from about 0.001 mg/kg to 7 mg/kg of body weight per day.

The nucleic acids of the present invention can be administered by known routes of administration including intravenous administration, intramuscular, intraperitoneal, intracerebrospinal, subcutaneous, intra-articular, intrasynovial, intrathecal, oral, topical, or inhalation routes. Effective dosages and schedules for administering antagonists or agonists are determined empirically according to guidelines generally recognized by those of skill in the art. Single or multiple dosages may be employed.

As noted above, the nucleic acids of the present invention useful in the methods of the present disclosure can be incorporated into pharmaceutical compositions suitable for administration into an animal such as a mammal. Methods for formulating such compositions are generally well known. Guidance is available for example from Remington: THE SCIENCE AND PRACTICE OF PHARMACY, 19th Edition, Gennaro (ed.) 1995, Mack Publishing Company, Easton, Pa. Such compositions typically comprise at least one anti-RT aptamer and a pharmaceutically acceptable carrier. The term "pharmaceutically acceptable carrier" refers to any and all coatings, excipients, solvents, dispersion media, absorption delaying agents, and the like, compatible with pharmaceutical administration. Such carriers also include for example sodium chloride, colloidal silica, talc, various polymeric carriers including polyvinyl pyrrolidone, cellulose-based compounds such as carboxymethylcellulose or methylcellulose, polyvinylpyrrolidone, polyacrylates, and polyethylene glycol. Dosage forms include, for example, oral or sublingual tablets, pellets, micro- and nano-capsules, liposomes, inhalation forms, nasal sprays, and sustained-release preparations. Solutions or suspensions used for administering nucleic acids of the present invention can include one or more of the following components: a sterile diluent such as water for injection, saline

solution; fixed oils, polyethylene glycols, glycerine, propylene glycol or other synthetic solvents; antibacterial agents such as benzyl alcohol or methyl parabens; antioxidants such as ascorbic acid or sodium bisulfite; chelating agents such as EDTA; buffers such as acetates, citrates or phosphates and agents for the adjustment of tonicity such as sodium chloride or dextrose. In some embodiments, a pharmaceutical composition can be delivered via slow release formulation or matrix comprising nucleic acids of the present invention or DNA constructs suitable for expression of nucleic acids of the present invention in or around a site within the body.

In some embodiments, the nucleic acid of the present invention of the invention may be formulated into pharmaceutical compositions that can be used to apply microbicides to effectively prevent transmission of HIV-1 through mucosae, more particularly to prevent the sexual or vaginal transmission of HIV-1. Thus, the compositions are in forms adapted to be applied to the site where sexual intercourse or related intimate contact takes place, such as the genitals, vagina, vulva, cervix, rectum, mouth, hands, lower abdomen, upper thighs, especially the vagina, vulva, cervix, and ano-rectal mucosae. As appropriate topical compositions there may be cited for example gels, jellies, creams, pastes, emulsions, dispersions, ointments, films, sponges, foams, aerosols, powders, intravaginal rings or other intravaginal drug delivery systems, cervical caps, implants, patches, suppositories or pessaries for rectal, or vaginal application, vaginal or rectal or buccal tablets, mouthwashes. The present topical formulations such as the gel formulations described herein could, for example, be applied into the vagina by hand, suppositories, or conventional tampon or syringe techniques. The method of administering or delivering the gel into the vagina is not critical so long as an effective amount of the gel is delivered into the vagina. The present topical formulations such as the gel formulations described herein may also be used for protection during anal intercourse and can be applied using similar techniques. For vaginal heterosexual intercourse, the present topical formulations such as the gel formulations described herein may be applied into the vagina prior to intercourse. For anal intercourse (heterosexual or homosexual), the present topical formulations such as the gel formulations described herein may be inserted into the rectum prior to intercourse. For either vaginal or anal intercourse, the present topical formulations such as the gel formulations described herein may also act as a lubricant. For added protection it is generally preferred that the present topical formulations such as the gel formulations described herein be applied before intercourse or other sexual activity and that, if appropriate, a condom be used. For even further protection, the present topical formulations

such as the gel formulations described herein can be applied as soon as possible after completion of the sexual activity. Although application only after the sexual activity is less recommended, it would still be desirable afterwards if the application was not performed prior to the sexual activity for any reason (e.g., in cases of rape).

In some embodiments, the nucleic acid of the present inventions of the invention may be used in all the suitable formulations, alone or in combination with other active ingredients, such as antivirals, antibiotics, immunomodulators or vaccines. They may also be used alone or in combination with other prophylactic agents for the prevention of viral infections. Thus, the nucleic acid of the present inventions of the invention may be combined with pharmaceutically acceptable adjuvants conventionally employed in vaccines and administered in prophylactically effective amounts to protect individuals over an extended period of time against HIV-1 infection. Antiviral compounds which may be used in combination with the nucleic acid of the present inventions of the invention may be known antiretroviral compounds such as pentamidine, thymopentin, castanospermine, dextran (dextran sulfate), foscarnet-sodium (trisodium phosphono formate); nucleoside reverse transcriptase inhibitors, e.g. zidovudine (3'-azido-3'-deoxythymidine, AZT), didanosine (2',3'-dideoxyinosine; ddI), zalcitabine (dideoxycytidine, ddC) or lamivudine (2'-3'-dideoxy-3'-thiacytidine, 3TC), stavudine (2',3'-didehydro-3'-deoxythymidine, d4T), abacavir and the like; non-nucleoside reverse transcriptase inhibitors such as nevirapine (11-cyclopropyl-5,11-dihydro-4-methyl-6H-dipyrido-[3,2-b:2',3'-e][1,4]diazepin-6-one), efavirenz, delavirdine, and the like; phosphonate reverse transcriptase inhibitors, e.g. tenofovir and the like; compounds of the TIBO (tetrahydro-imidazo[4,5,1-jk][1,4]-benzodiazepine-2(1H)-one and thione)-type e.g. (S)-8-chloro-4,5,6,7-tetrahydro-5-methyl-6-(3-methyl-2-butenyl)imidazo-[4,5,1-jk][1,4]benzo-diazepine-2(1H)-thione; compounds of the [alpha]-APA ([alpha]-anilino phenyl acetamide) type e.g. [alpha]-[(2-nitrophenyl)amino]-2,6-dichlorobenzene-acetamide and the like; inhibitors of trans-activating proteins, such as TAT-inhibitors, e.g. RO-5-3335, or REV inhibitors, and the like; protease inhibitors e.g. indinavir, ritonavir, saquinavir, lopinavir (ABT-378), nelfinavir, amprenavir, TMC-126, BMS-232632, VX-175 and the like; fusion inhibitors, e.g. T-20, T-1249 and the like; CXCR4 receptor antagonists, e.g. AMD-3100 and the like; inhibitors of the viral integrase; ribonucleotide reductase inhibitors, e.g. hydroxyurea and the like. Combinations may as well exert a synergistic effect in inhibiting HIV-1 replication when components of the combination act on different or same sites of HIV-1 replication, preferably on different sites. The use of such combinations may reduce the dosage

of a given conventional antiretroviral agent which would be required for a desired prophylactic effect as compared to when that agent is administered as a single active ingredient. These combinations reduce potential of resistance to single agent, while minimizing any associated toxicity. These combinations may also increase the efficacy of the conventional agent without increasing the associated toxicity.

The invention will be further illustrated by the following figures and examples. However, these examples and figures should not be interpreted in any way as limiting the scope of the present invention.

FIGURES:

Figure 1: A. genetic structure of the HIV-1 genome. B. Bioinformatic search of G4 forming sequences. This graphical representation show the score in function of the oligonucleotidic sequence of the HIV genome.

Figure 2: Examples of UV-melting profiles recorded at 295 nm at a strand concentration of 5 µM for CPPT (ID=11). The experiments were performed in a buffer composed of 20 mM potassium phosphate pH 6.9 supplemented with 70 mM KCl.

Figure 3: HeLa P4 cells were infected by viral supernatant HIV-1˙ as described in the Methods section. To assess the inhibitory effect of the decoys, HeLa P4 cells were infected with HIV-1 in the presence of various concentrations of oligonucleotide. Viral replication was monitored by Beta-galactosidase activity at 24h post-infection. Values were normalized to 100% corresponding to Beta-galactosidase activity in the absence of decoys. A, B, C HIV-1 replication percentage measured in the presence of increasing concentrations of decoys from 0.1 nM to 500 nM of decoys. The data were collected for the following sequences: A) PRO1 (ID=8, black squares), PRO2 (ID=6, black circles), PRO3 (ID=4, grey squares), PRO4 (ID=2, black lozenge). B) CPPT (ID=11, black squares), PPT (ID=10, black circles), 93del (grey squares), C) rPRO1 (ID=9, black squares), rPPT (ID=13, black circles), rCPPT (ID=12, grey squares).

EXAMPLE:

Material & Methods

Bioinformatics analysis: The 1870 HIV-1 sequences were obtained from the HIV-1 database which provides premade alignments to the community (www.hiv.lanl.gov). The alignment apparently presents around 11 000 nucleotides instead of the usual 9200 nucleotides for the HIV-1 consensus sequence. This is due to the insertion of gaps to optimise the alignment of the 1870 sequences. The score algorithm that we developed in the laboratory (Bedrat et al, in preparation) searches for G/C skewness and the presence of GC blocks in the alignment of the 1870 HIV-1 sequences retrieved from the HIV-1 database. It analyses the genome using a sliding window of 25 nucleotides and attribute a score to the first nucleotide of the window. The average of the 1870 scores obtained for each window is depicted in a graphical representation. We consider that the scores higher than 1 in absolute value (-1 or +1) are potentially able to form DNA or RNA G4 structures in the (+) strand (for positive values) or in the (-) strand for the negative values. The analysis of sequence conservation was performed using the webLOGO software to generate the LOGO representation.

Preparation of the oligonucleotides : Oligonucleotides were purchased from Eurogentec (Seraing, Belgium) without further purification (Reverse-Phase Cartridge Gold). Concentrations were determined by ultraviolet (UV) absorption using the extinction coefficients provided by the manufacturer. All oligonucleotides were dissolved in 20 mM potassium phosphate buffer containing 70 mM KCl.

UV-Melting: UV-melting measurements[8] were performed on a Uvikon XL (Secomam) spectrophotometer coupled to a water bath temperature-control accessory. A temperature-increase rate of 0.2°C/min was applied and the absorbance values were measured every 1°C. The temperature was measured with an inert glass sensor immersed into a control quartz cell filled with water. The absorbance was monitored at 240 and 295 nm using quartz cells of 0.2 or 1 cm pathlength and 580 µl of volume.

Cell lines and viruses : HeLa P4 cells encoding a Tat-inducible β-galactosidase were maintained in DMEM medium (Invitrogen) supplemented with 10 % inactivated FCS, 1 mg/ml geneticin (G418, Gibco-BRL), gentamycin. MT4 and H9Laï cells were grown in RPMI 1640 glutamax medium (Invitrogen) supplemented with 10 % inactivated FCS. HIV-1 viruses were obtained after 48 h co-culture of MT4 cells (0,5 x10^6 /ml) and H9Laï cells (1 x

10^6 /ml), chronically infected by HIV-1Laï isolate, in RPMI 1640 glutamax medium supplemented with 10 % inactivated FCS, at 37°C under humidified atmosphere and 5 % CO2. The culture was then centrifuged and the supernatant was clarified by filtration on a 0.45 µm membrane before freezing at -80°C.

Viral infectivity : The oligonucleotides were preincubated in a 100 mM potassium solution to favour G4 formation. When added they are incubated in presence of the HelaP4 cells 20 minutes before infection. The infectivity was assayed on HeLa P4 cells expressing CD4 receptor and the β-galactosidase gene under the control of the HIV-1 LTR. HeLa P4 were plated using 200 µl of DMEM medium supplemented with 10 % inactivated FCS in 96-multi-well plates at 10 000 cells per well. After overnight incubation at 37°C, under humidified atmosphere and 5 % CO2, the supernatant was discarded and 200 µl of viral preparation were added in serial dilutions. After 24 h of infection, the supernatant was discarded and the wells were washed 3 times with 200 µl of PBS. Each well was refilled with 200 µl of a reaction buffer containing 50 mM Tris-HCl pH 8,5, 100 mM β- mercaptoethanol, 0,05 % Triton X-100 and 5 mM 4-methylumbelliferyl-B-D-galactopyranoside (4-MUG) (Sigma). After 24 h, the reaction was measured in a fluorescence microplate reader (Cytofluor II, Applied Biosystems) at 360/460 nm Ex/Em.

Results

Bioinformatics search of G4 forming sequences in the HIV-1 genome: Using our bioinformatic algorithm we searched for sequences that are potentially able to form G4 structures in the HIV-1 genome. We analysed an alignment of 1870 viral sequences provided by the HIV-1 database. We found 10 different loci that could potentially form G4 structures (Figure 1). Nine of them are present in the (+) strand with scores higher than +1 and one is present in the (-) strand with a score < -1.

Verification of the G4 formation *in vitro*: To verify the formation of G4 structures, we purchased the DNA oligonucleotide corresponding to these sequences and we performed UV-melting experiments followed at 295 nm (Figure 2). At this wavelength the denaturation of the structure generates an hypochromism that is specific of the G4 structures. Amongst the 10 detected candidates, sequence # 1, 2 and 4 formed very unstable G4 structures, with melting temperature below 10°C (data not shown), that might not be compatible with in vivo

formation. The seven remaining sequences formed stable RNA and DNA G4 structures with melting temperature ranging from 30°C to 75°C. In figure 2A are presented exemples of G4 specific melting profiles. In table 1 are presented the melting temperatures derived from the melting profiles of the 3 most important sequences for this study. We decided to analyse the sequence # 10 (45 bases) by synthesizing smaller tracts of 19 to 23 nucleotides spanning the entire sequence. This sequence is therefore able to form 4 different G4 structures (DNA or RNA backbones) with melting temperature ranging from 40°C to 75°C. We also determined the G4 structure[9] of the PR02 sequence (ID=6) by Nuclear Magnetic Resonance spectroscopy. It forms a stable two G-tetrads antiparallel G4 with an additional Watson-Crick CG base pair. We also confirmed that sequence # 3 (CPPT (ID=11)) and # 9 (PPT (ID=10)) are also able to form stable G4 structures with melting temperature of 59°C and 35°C.

N°	Name	**SEQ ID**	**SEQUENCES**	**T_m (°C)**	**Inhibition IC_{50} (nM)**
10	Pro	**1**	GGGACTTTCCGGGGAGGCGTGGCCTGGGCGGGACTGGGGAGTGGC	-	-
10-1	Pro4	**2**	GGGACTTTCCGGGGAGGTGTGGC	40	5.5
	rPro4	**3**	GGGACUUUCCGGGGAGGUGUGGC (RNA)	53	200
10-2	Pro3	4	AGGGAGGCGTGGCCTGGGCGGG	58	45
	rPro3	5	AGGGAGGCGUGGCCUGGGCGGG (RNA)	70	< 100
10-3	Pro2	6	TGGCCTGGGCGGGACTGGG	56	100
	rPro2	7	UGGCCUGGGCGGGACUGGG (RNA)	75	~ 10
10-4	Pro1	8	GGGCGGGACTGGGGAGTGGC	58	1.4
	rPro1	9	GGGCGGGACUGGGGAGUGGC	69	4
9	PPT	10	AAGGGGGGACTGGATGGGCT	35	6.5
3	CPPT	11	AAGGGGGGATTGGGGGGTACAGTGCAGGGGGAA	59	0.9
3	rCPPT	12	AAGGGGGGAUUGGGGGGUACAGUGCAGGGGGAA (RNA)	65	3
9	rPPT	13	AAGGGGGGACUGGAUGGGCU (RNA)	45	3.5

Potential biological roles of these sequences: According to their locations, we also found that these three sequences might play a role in key steps of the viral cycle: i) Sequence # 3 is localised in the center of the viral genome and it contains the C-ppt sequence. This important sequence is the central initiation site of the reverse transcription during the (+) strand DNA synthesis. ii) Sequence # 9 is localised at 3' end of the genome. These sequence is the first initiation site of the reverse transcription during the (+) strand DNA synthesis. iii) Sequence # 10 is located on the promoter of the provirus, at 40 nt upstream from the transcription initiation site, close to the TATA box. This sequence overlaps the so-called minimum promoter composed of three SP1 and two NF-kB binding sites which are crucial for

the initiation of the transcription of HIV-1. The LOGO representation generated from the alignment of the 1870 sequences shows a high level of conservation of the 3 sequences suggesting an important role involving specific protein recognition of these sequences.

Inhibition of the viral infectivity using a decoy strategy: These data prompted us to test if these oligonucleotides are able to inhibit HIV-1 infectivity as already observed for Andevir or Zintevir. We purchased and tested these sequences in a viral infectivity test realised in vivo with real HIV viruses infecting HeLap4 cells. The G4s derived from HIV-1 genome strongly inhibited HIV-1 infectivity with IC_{50} lower that 4 nM for Sequences # 3 and 10-4 (Figure 3). The inhibition observed for the decoys are much stronger than the ones observed for Andevir (93del) (IC_{50} = 25nM). Cytotoxicity test on HeLaP4 cells performed with the same G4s at a concentration of 500 nM did not reveal any toxicity after 24 hours. These G4s might therefore act as decoys and may trap crucial proteins involved in the recognition of the same sequences present in the HIV-1 genome.

REFERENCES:

Throughout this application, various references describe the state of the art to which this invention pertains. The disclosures of these references are hereby incorporated by reference into the present disclosure.

(1) Pomerantz, R. J.; Horn, D. L. *Nat Med* **2003**, *9*, 867.

(2) Webba da Silva, M.; Trajkovski, M.; Sannohe, Y.; Ma'ani Hessari, N.; Sugiyama, H.; Plavec, J. *Angew Chem Int Ed Engl* **2009**, *48*, 9167.

(3) Biffi, G.; Di Antonio, M.; Tannahill, D.; Balasubramanian, S. *Nat Chem* **2014**, *6*, 75.

(4) Biffi, G.; Tannahill, D.; McCafferty, J.; Balasubramanian, S. *Nat Chem* **2013**, *5*, 182.

(5) Murat, P.; Zhong, J.; Lekieffre, L.; Cowieson, N. P.; Clancy, J. L.; Preiss, T.; Balasubramanian, S.; Khanna, R.; Tellam, J. *Nat Chem Biol* **2014**, *10*, 358.

(6) Faure-Perraud, A.; Metifiot, M.; Reigadas, S.; Recordon-Pinson, P.; Parissi, V.; Ventura, M.; Andreola, M. L. *Antivir Ther* **2011**, *16*, 383.

(7) de Soultrait, V. R.; Lozach, P. Y.; Altmeyer, R.; Tarrago-Litvak, L.; Litvak, S.; Andreola, M. L. *J Mol Biol* **2002**, *324*, 195.

(8) Mergny, J. L.; Lacroix, L. *Curr Protoc Nucleic Acid Chem* **2009**, *Chapter 17*, Unit 17 1.

(9) Amrane, S.; Kerkour, A.; Bedrat, A.; Vialet, B.; Andreola, M. L.; Mergny, J. L. *J Am Chem Soc* **2014**, *136*, 5249.

5

CLAIMS:

1. A nucleic acid (NA3) having the general formula (I) of :

   $L3_1$-$(G)_{4-6}$-$L3_2$-$(G)_{4-6}$-$L3_3$-$(G)_{3-4}$-$L3_4$ (I)

   wherein

   - (G)n represents a sequence of n guanosines
   - $L3_1$ represents a sequence of 0 to 4 nucleotides
   - $L3_2$ represents a sequence of 2 to 5 nucleotides
   - $L3_3$ represents a sequence of 5 to 10 nucleotides
   - $L3_4$ represents a sequence of 0 to 4 nucleotides

2. The nucleic acid of claim 1 wherein $L3_1$ represents AA.

3. The nucleic acid of claim 1 wherein $L3_2$ represents ATT or AUU.

4. The nucleic acid of claim 1 wherein $L3_3$ represents TACAGTGCA or UACAGUGCA.

5. The nucleic acid of claim 1 wherein $L3_4$ represents AA.

6. The nucleic acid of claim 1 which is represented by SEQ ID NO: 11 or SEQ ID NO:12.

7. A nucleic acid (NA9) having the general formula (II) of:

   $L9_1$-$(G)_{4-6}$-$L9_2$-$(G)_{2-3}$-$L9_3$-$(G)_{2-4}$-$L9_4$ (II)

   Wherein

   - (G)n represents a sequence of n guanosines
   - $L9_1$ represents a sequence of 0 to 4 nucleotides
   - $L9_2$ represents a sequence of 2 to 4 nucleotides
   - $L9_3$ represents a sequence of 1 to 4 nucleotides

- $L9_4$ represents a sequence of 0 to 4

8. The nucleic acid of claim 7 wherein $L9_1$ represents AA.

9. The nucleic acid of claim 7 wherein $L9_2$ represents ACT or ACU.

10. The nucleic acid of claim 7 wherein $L9_3$ represents AT or AU.

11. The nucleic acid of claim 7 wherein $L9_4$ represents AA.

12. The nucleic acid of claim 7 which is represented by SEQ ID NO:10 or SEQ ID NO:13.

13. A nucleic acid (NA10) having the general formula (III) of:

$L10_0$-$(G)_{3-4}$-$L10_1$-$(G)_{3-4}$-$L10_2$-$(G)_{2-3}$-$L10_3$-$(G)_{2-3}$-$L10_4$-$(G)_{3-4}$-$L10_5$-$(G)_{3-4}$-$L10_6$-$(G)_{3-4}$-$L10_7$- $(G)_{2-3}$-$L10_8$ (III)

Wherein

- (G)n represents a sequence of n guanosines
- $L10_0$ represents a sequence of 0 to 4 nucleotides
- $L10_1$ represents a sequence of 7 to 10 nucleotides
- $L10_2$ represents a sequence 1 to 3 nucleotides
- $L10_3$ represents a sequence of 1 to 4 nucleotides
- $L10_4$ represents a sequence 2 to 4 nucleotides
- $L10_5$ represents a single nucleotide
- $L10_6$ represents a sequence of 2 to 5 nucleotides
- $L10_7$ represents a sequence of 2 to 5 nucleotides
- $L10_8$ represents a sequence of 0 to 4 nucleotides

14. The nucleic acid of claim 13 wherein $L10_1$ represents ACTTTCC or ACUUUCC.

15. The nucleic acid of claim 13 wherein $L10_2$ represents A.

16. The nucleic acid of claim 13 wherein $L10_3$ represents CGT or CGU.

17. The nucleic acid of claim 13 wherein $L10_4$ represents CCT or CCU.

18. The nucleic acid of claim 13 wherein $L10_5$ represents C.

19. The nucleic acid of claim 13 wherein $L10_6$ represents ACT or ACU.

20. The nucleic acid of claim 13 wherein $L10_7$ represents AGT or AGU.

21. The nucleic acid of claim 13 which is represented by SEQ ID NO:1.

22. A nucleic acid (NA10.1) having the general formula (IV) of:

$L10_0$-$(G)_{3-4}$-$L10_1$-$(G)_{3-4}$-$L10_2$-$(G)_{2-3}$-$L10_3$-$(G)_{2-3}$-$L10_4$ (IV)

Wherein

- (G)n represents a sequence of n guanosines
- $L10_0$ represents a sequence of 0 to 4 nucleotides
- $L10_1$ represents a sequence of 7 to 10 nucleotides
- $L10_2$ represents a sequence of 1 to 3 nucleotides
- $L10_3$ represents a sequence of 1 to 4 nucleotides
- $L10_4$ represents a sequence of 0 to 4 nucleotides

23. The nucleic acid of claim 22 wherein $L10_1$ represents ACTTTCC or ACUUUCC.

24. The nucleic acid of claim 22 wherein $L10_2$ represents A.

25. The nucleic acid of claim 22 wherein $L10_3$ represents CGT or CGU.

26. The nucleic acid of claim 22 wherein $L10_4$ represents C.

27. The nucleic acid of claim 22 which is represented by SEQ ID NO:2 or SEQ ID NO:3.

28. A nucleic acid (NA10.2) having the general formula (V) of:

$L10_1$-$(G)_{3-4}$-$L10_2$-$(G)_{2-3}$-$L10_3$-$(G)_{2-3}$-$L10_4$-$(G)_{3-4}$-$L10_5$-$(G)_{3-4}$-$L10_6$ (V)

Wherein

- $L10_1$ represents a sequence of 0 to 4 nucleotides
- $L10_2$ represents a sequence of 1 to 3 nucleotides
- $L10_3$ represents a sequence of 1 to 4 nucleotides
- $L10_4$ represents a sequence of 2 to 4 nucleotides
- $L10_5$ represents a single nucleotide
- $L10_6$ represents a sequence of 0 to 4 nucleotides

29. The nucleic acid of claim 28 wherein $L10_1$ represents A.

30. The nucleic acid of claim 28 wherein $L10_2$ represents A.

31. The nucleic acid of claim 28 wherein $L10_3$ represents CGT or CGU.

32. The nucleic acid of claim 28 wherein $L10_4$ represents CCT or CCU.

33. The nucleic acid of claim 28 wherein $L10_5$ represents C.

34. The nucleic acid of claim 28 which is represented by SEQ ID NO:4 or SEQ ID NO:5.

35. A nucleic acid (NA10.3) having the general formula (VI) of:

$L10_3$-$(G)_{2-3}$-$L10_4$-$(G)_{3-4}$-$L10_5$-$(G)_{3-4}$-$L10_6$-$(G)_{3-4}$-$L10_7$ (VI)

Wherein

- (G)n represents sequence of n guanosines
- $L10_3$ represents a sequence of 0 to 4 nucleotides
- $L10_4$ represents a sequence 2 to 4 nucleotides

- $L10_5$ represents a single nucleotide
- $L10_6$ represents a sequence of 2 to 5 nucleotides
- $L10_7$ represents a sequence of 0 to 4 nucleotides

36. The nucleic acid of claim 35 wherein $L10_3$ represents T or U.

37. The nucleic acid of claim 35 wherein $L10_4$ represents CCT or CCU.

38. The nucleic acid of claim 35 wherein $L10_5$ represents C.

39. The nucleic acid of claim 35 wherein $L10_6$ represents ACT or ACU.

40. The nucleic acid of claim 35 which is represented by SEQ ID NO:6 or SEQ ID NO:7.

41. A nucleic acid (NA10.4) having the general formula (VII) of:

$L10_4$-$(G)_{3-4}$-$L10_5$-$(G)_{3-4}$-$L10_6$-$(G)_{3-4}$-$L10_7$- $(G)_{2-3}$-$L10_8$ (VII)

Wherein

- (G)n represents a sequence of n guanosines
- $L10_4$ represents a sequence of 0 to 4 nucleotides
- $L10_5$ represents a single nucleotide
- $L10_6$ represents a sequence of 2 to 5 nucleotides
- $L10_7$ represents a sequence of 2 to 5 nucleotides
- $L10_8$ represents a sequence of 0 to 4 nucleotides

42. The nucleic acid of claim 41 wherein $L10_4$ represents CCT or CCU.

43. The nucleic acid of claim 41 wherein $L10_5$ represents C.

44. The nucleic acid of claim 41 wherein $L10_6$ represents ACT or ACU.

45. The nucleic acid of claim 41 wherein $L10_7$ represents AGT or AGU.

46. The nucleic acid of claim 41 which is represented by SEQ ID NO:8 or SEQ ID NO:9.

47. A method of treating a lentivirus infection in a subject in need thereof comprising administering the subject with a therapeutically effective amount of at least one nucleic acid according to any one of claims 1 to 46.

48. The nucleic acid according to any one of claims 1 to 46 for use as a medicament.

49. A pharmaceutical composition comprising a nucleic acid according to any one of claims 1 to 46.

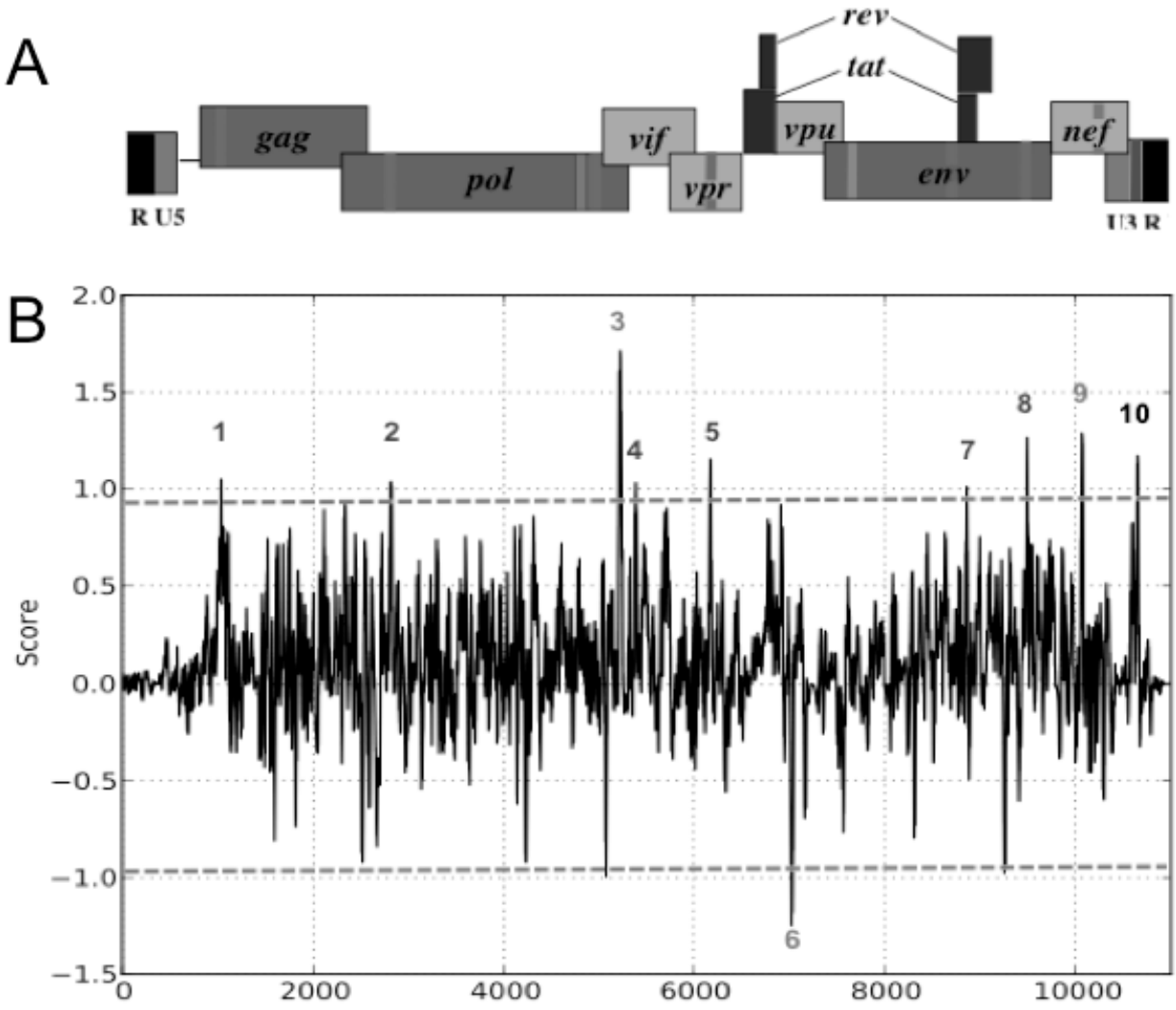

A
rev
tat
gag
pol
vif
vpr
vpu
env
nef
R U5
U3 R
B
Score
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
0
2000
4000
6000
8000
10000
1
2
3
4
5
6
7
8
9
10

Figure 1

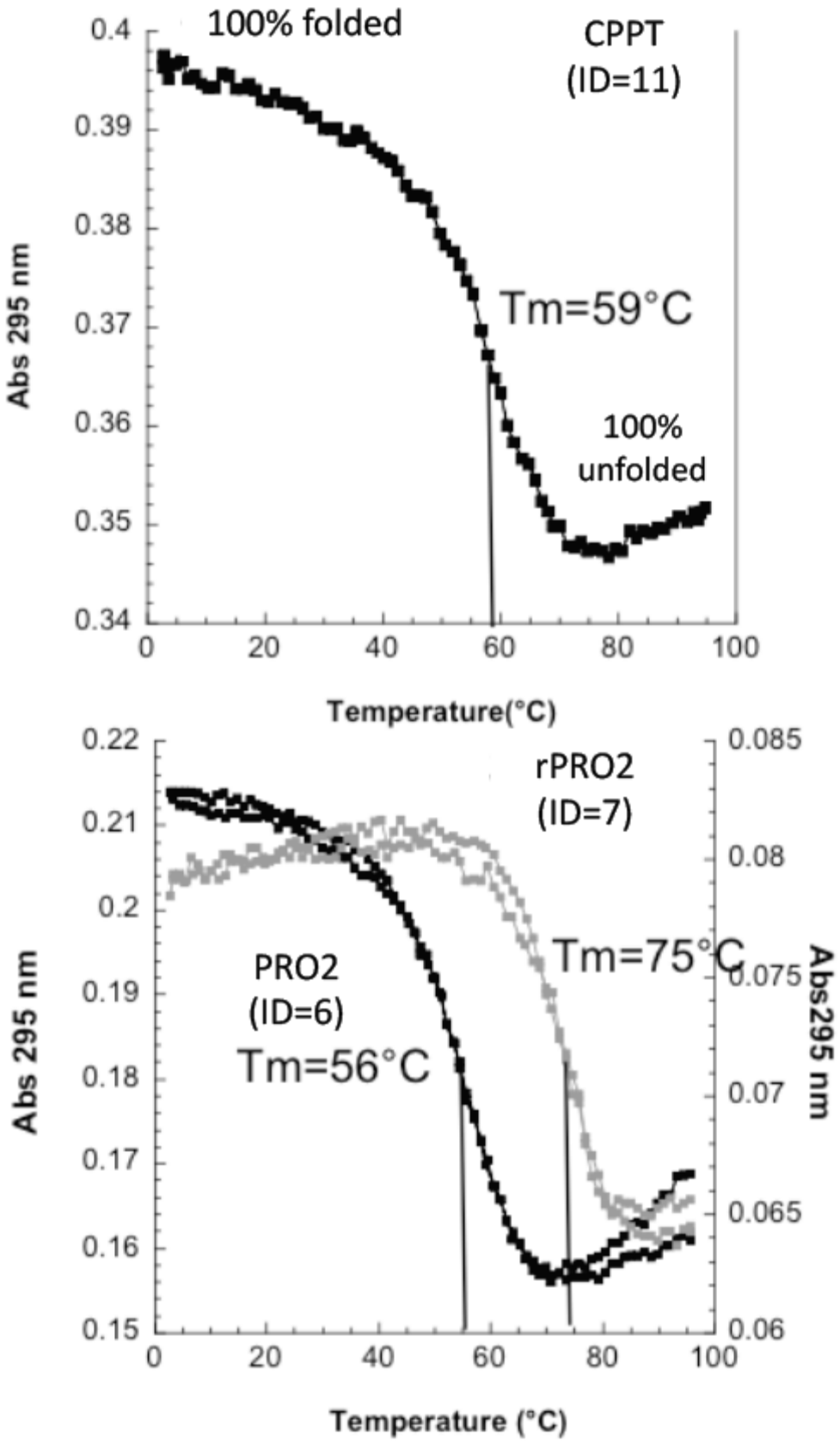

100% folded
CPPT
(ID=11)
Tm=59°C
100%
unfolded
Abs 295 nm
Temperature(°C)
rPRO2
(ID=7)
Tm=75°C
PRO2
(ID=6)
Tm=56°C
Abs 295 nm
Abs295 nm
Temperature (°C)

Figure 2

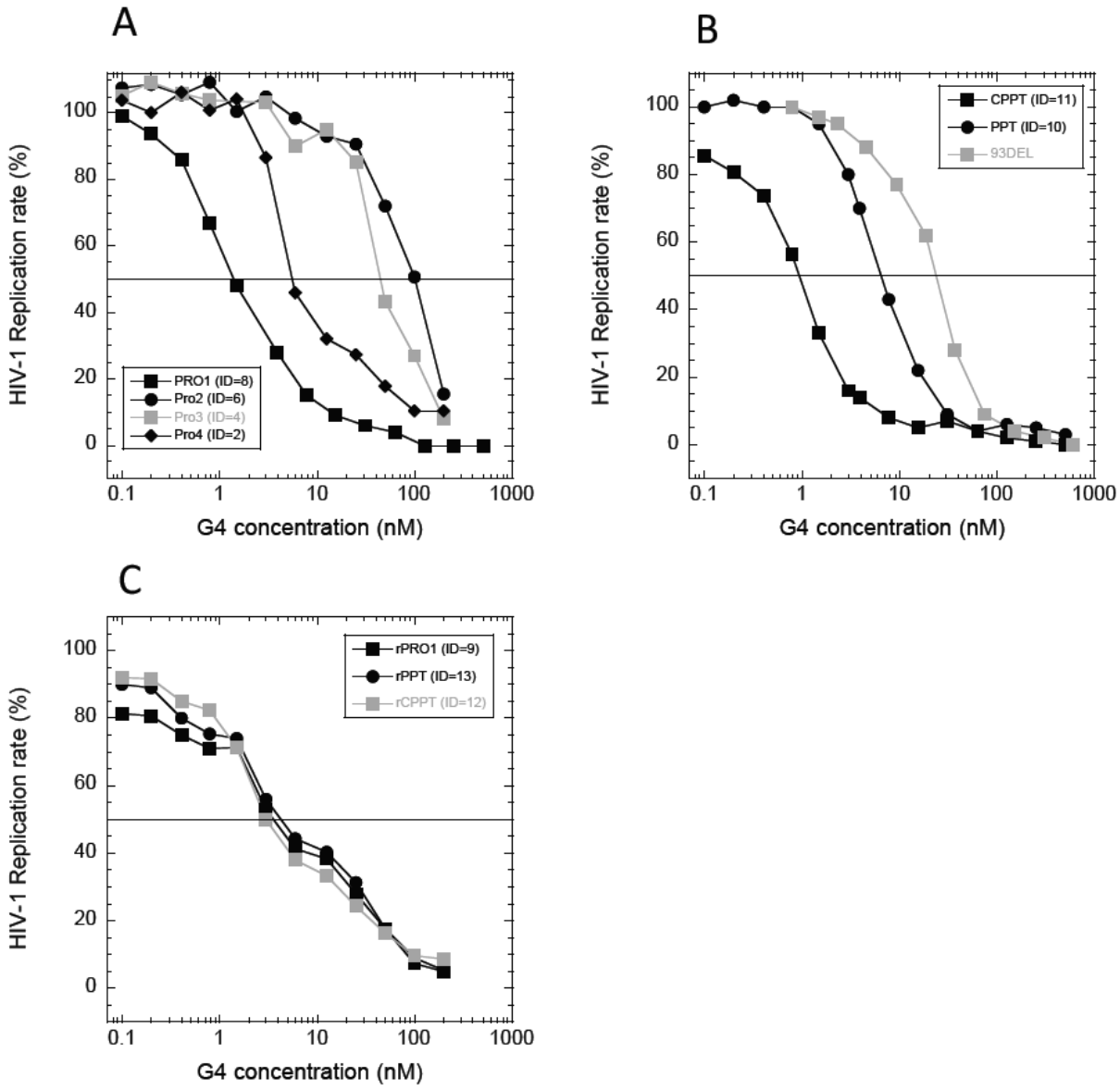

A
HIV-1 Replication rate (%)
G4 concentration (nM)
PRO1 (ID=8)
Pro2 (ID=6)
Pro3 (ID=4)
Pro4 (ID=2)
B
HIV-1 Replication rate (%)
G4 concentration (nM)
CPPT (ID=11)
PPT (ID=10)
93DEL
C
HIV-1 Replication rate (%)
G4 concentration (nM)
rPRO1 (ID=9)
rPPT (ID=13)
rCPPT (ID=12)

Figure 3

5.5 Patent 2: Methods and pharmaceutical compositions for the treatment of filovirus infections.

A. Bedrat[1,2], S.Amrane[1,2], & JL Mergny[1,2]

EP 15 305 367.3, deposit Mai 24^{th} 2015.

1. Université de Bordeaux, ARNA Laboratory, F-33000 Bordeaux, France;
2. Inserm U869, IECB, F-33600 Pessac, France;

The present invention relates to methods and pharmaceutical compositions for the treatment of filovirus infections. In particular, the present invention relates to a method of treating filovirus infection in a subject in need thereof comprising administering the subject with a therapeutically effective amount of at least one oligonucleotide comprising the sequence as set forth in SEQ ID NO:1 to SEQ ID NO:15.

METHODS AND PHARMACEUTICAL COMPOSITIONS FOR THE TREATMENT OF FILOVIRUS INFECTIONS

FIELD OF THE INVENTION:

The present invention relates to methods and pharmaceutical compositions for the treatment of filovirus infections.

BACKGROUND OF THE INVENTION:

The family Filoviridae (Filovirus) is the taxonomic home of several related viruses that form filamentous infectious viral particles (virions), and encode their genome in the form of single-stranded negative-sense RNA. Two members of the family that are commonly known are Ebola virus and Marburg virus. Both viruses, and some of their lesser known relatives, cause severe disease in humans and nonhuman primates in the form of viral hemorrhagic fevers. The Ebola virus was named after the Ebola River in Zaire, Africa, near where the first outbreak was noted by Dr. Ngoy Mushola in 1976 after a significant outbreaks in both Yambuku, Zaire (now the Democratic Republic of the Congo), and Nzara, in western Sudan. There are three distinct species of Ebola virus which cause fatal disease in humans: Zaire ebolavirus (ZEBOV) (also known as EBOV), Sudan ebolavirus (SEBOV) and Ivory Coast ebolavirus (ICEBOV). Among humans, the Ebola virus is transmitted by direct contact with infected body fluids such as blood. The incubation period of Ebola virus infection varies from two days to four weeks. Symptoms are variable too, but the onset is usually sudden and characterised by high fever, prostration, myalgia, arthralgia, abdominal pains and headache. These symptoms progress to vomiting, diarrhea, oropharyngeal lesions, conjunctivitis, organ damage (notably the kidney and liver) by co-localized necrosis, proteinuria, and bleeding both internal and external, commonly through the gastrointestinal tract. Death or recovery to convalescence occurs within six to ten days of onset of symptomology. Although several antivirals have shown efficacy against Ebola virus infection *in vitro* or in animal models, few of them have been yet assessed in human beings with Ebola virus disease. Thus, there exists a huge need in the art for an effective curative treatment against Ebola Virus Disease.

SUMMARY OF THE INVENTION:

The present invention relates to methods and pharmaceutical compositions for the treatment of filovirus infections. In particular, the present invention is defined by the claims.

DETAILED DESCRIPTION OF THE INVENTION:

The present invention relates to a method of treating filovirus infection in a subject in need thereof comprising administering the subject with a therapeutically effective amount of at least one oligonucleotide comprising the sequence as set forth in SEQ ID NO:1 to SEQ ID NO:15.

The term “filovirus” refers collectively to members of the Filoviridae family of single stranded (−) RNA viruses including Ebola and Marburg viruses. As used herein, the term "Ebola virus" refers to a member of the family Filoviridae, are associated with outbreaks of highly lethal hemorrhagic fever in humans and nonhuman primates. Human pathogens include Ebola Zaire, Ebola Sudan, and Ebola Ivory Coast. Ebola Reston is a monkey pathogen and is not considered a significant human pathogen. In some embodiments of the invention, said Ebola virus is Ivory Coast Ebola virus (ICEBOV), Zaire Ebola virus (ZEBOV or EBOV), Sudan Ebola Virus (SEBOV), or a new strain or species of Ebola virus.

The method of the present invention is particularly suitable for the treatment of filovirus diseases, in particular Ebola virus disease. As used herein, the term "Ebola virus disease" (EVD), formerly known as Ebola haemorrhagic fever, is a severe, often fatal illness in humans. The incubation period, that is, the time interval from infection with the virus to onset of symptoms is 2 to 21 days. Humans are not infectious until they develop symptoms. First symptoms are the sudden onset of fever fatigue, muscle pain, headache and sore throat. This is followed by vomiting, diarrhoea, rash, symptoms of impaired kidney and liver function, and in some cases, both internal and external bleeding (e.g. oozing from the gums, blood in the stools). Laboratory findings include low white blood cell and platelet counts and elevated liver enzymes.

As used herein, the term "treatment" or "treat" refer to both prophylactic or preventive treatment as well as curative or disease modifying treatment, including treatment of subjects at risk of contracting the disease or suspected to have contracted the disease as well as subjects who are ill or have been diagnosed as suffering from a disease or medical condition, and includes suppression of clinical relapse. The treatment may be administered to a subject having a medical disorder or who ultimately may acquire the disorder, in order to prevent,

cure, delay the onset of, reduce the severity of, or ameliorate one or more symptoms of a disorder or recurring disorder, or in order to prolong the survival of a subject beyond that expected in the absence of such treatment. By "therapeutic regimen" is meant the pattern of treatment of an illness, e.g., the pattern of dosing used during therapy. A therapeutic regimen may include an induction regimen and a maintenance regimen. The phrase "induction regimen" or "induction period" refers to a therapeutic regimen (or the portion of a therapeutic regimen) that is used for the initial treatment of a disease. The general goal of an induction regimen is to provide a high level of drug to a subject during the initial period of a treatment regimen. An induction regimen may employ (in part or in whole) a "loading regimen", which may include administering a greater dose of the drug than a physician would employ during a maintenance regimen, administering a drug more frequently than a physician would administer the drug during a maintenance regimen, or both. The phrase "maintenance regimen" or "maintenance period" refers to a therapeutic regimen (or the portion of a therapeutic regimen) that is used for the maintenance of a subject during treatment of an illness, e.g., to keep the subject in remission for long periods of time (months or years). A maintenance regimen may employ continuous therapy (e.g., administering a drug at a regular intervals, e.g., weekly, monthly, yearly, etc.) or intermittent therapy (e.g., interrupted treatment, intermittent treatment, treatment at relapse, or treatment upon achievement of a particular predetermined criteria [e.g., disease manifestation, etc.]).

In some embodiments, the subject is infected, or is at risk of being infected with a filovirus. Diagnosis may be performed by any suitable means. One skilled in the art will understand that a subject to be treated according to the present invention may have been identified using standard tests or may have been identified, without examination, as one at high risk due to the presence of one or more risk factors (e.g., exposure to Ebola virus, etc.). In some embodiments, the subject is infected but is asymptomatic (i.e. the symptoms are not detected). In some embodiments, the diagnosis is performed by detecting filovirus virus nucleic acids in a sample obtained from the subject by any method familiar to one of skill in the art. Such methods typically include the methods based on the detecting the filovirus virus nucleic acids expression. Filovirus nucleic acids may be detected in a RNA sample, preferably after amplification. For instance, the isolated RNA may be subjected to coupled reverse transcription and amplification, such as reverse transcription and amplification by polymerase chain reaction (RT-PCR), using specific oligonucleotide primers that are specific for a filovirus nucleic acid (e.g. those encoding the nucleoprotein (NP) and the four virion

structural proteins (VP40, VP35, VP30, and VP24). For instance, a RT-PCR Assay is intended for the *in vitro* qualitative detection of filovirus RNA in clinical specimens, including whole blood, serum, plasma, and urine, from individuals meeting filovirus clinical and/or epidemiological criteria (for example, clinical signs and symptoms associated with filovirus, contact with a probable or confirmed filovirus case, history of travel to geographic locations where filovirus cases were detected, or other epidemiologic links for which filovirus testing may be indicated as part of a public health investigation).

In some embodiments, the present invention contemplates the use of AS1411 (as described in WO2009098464) which has the sequence 5'-GGTGGTGGTGGTTGTGGTGGTGGTGG-3' (SEQ ID NO: 1) and is also known as GR026B and AGRO100. AS 1411 is a 26-mer DNA aptamer with unmodified phosphodiester linkages and forms a G-quadruplex structure (Dapic, V. et al. 2003) that is resistant to degradation by serum enzymes (Dapic, V. et al. 2002).

A further aspect of the present invention relates to an oligonucleotide comprising the sequence as set forth in SEQ ID NO:2 to SEQ ID NO:15.

For use in the instant invention, the oligonucleotide of the present invention is synthesized de novo using any of a number of procedures well known in the art. Chemical synthesis can be performed by a variety of automated nucleic acid synthesizers available in the market. These nucleic acids may be referred to as synthetic nucleic acids. Alternatively, the oligonucleotide of the present invention can be produced on a large scale in plasmids. The oligonucleotide of the present invention can be prepared from existing nucleic acid sequences using known techniques, such as those employing restriction enzymes, exonucleases or endonucleases.

By a "therapeutically effective amount" is meant a sufficient amount of the oligonucleotide of the present invention to treat the Filovirus infection at a reasonable benefit/risk ratio applicable to any medical treatment. It will be understood that the total daily usage of the compounds and compositions of the present invention will be decided by the attending physician within the scope of sound medical judgment. The specific therapeutically effective dose level for any particular subject will depend upon a variety of factors including the disorder being treated and the severity of the disorder; activity of the specific compound

employed; the specific composition employed, the age, body weight, general health, sex and diet of the subject; the time of administration, route of administration, and rate of excretion of the specific compound employed; the duration of the treatment; drugs used in combination or coincidental with the specific polypeptide employed; and like factors well known in the medical arts. For example, it is well within the skill of the art to start doses of the compound at levels lower than those required to achieve the desired therapeutic effect and to gradually increase the dosage until the desired effect is achieved. However, the daily dosage of the products may be varied over a wide range from 0.01 to 1,000 mg per adult per day. Preferably, the compositions contain 0.01, 0.05, 0.1, 0.5, 1.0, 2.5, 5.0, 10.0, 15.0, 25.0, 50.0, 100, 250 and 500 mg of the active ingredient for the symptomatic adjustment of the dosage to the subject to be treated. A medicament typically contains from about 0.01 mg to about 500 mg of the active ingredient, preferably from 1 mg to about 100 mg of the active ingredient. An effective amount of the drug is ordinarily supplied at a dosage level from 0.0002 mg/kg to about 20 mg/kg of body weight per day, especially from about 0.001 mg/kg to 7 mg/kg of body weight per day.

The oligonucleotide of the present invention can be administered by known routes of administration including intravenous administration, intramuscular, intraperitoneal, intracerebrospinal, subcutaneous, intra-articular, intrasynovial, intrathecal, oral, topical, or inhalation routes. Effective dosages and schedules for administering antagonists or agonists are determined empirically according to guidelines generally recognized by those of skill in the art. Single or multiple dosages may be employed.

As noted above, the oligonucleotide of the present invention useful in the methods of the present disclosure can be incorporated into pharmaceutical compositions suitable for administration into an animal such as a mammal. Methods for formulating such compositions are generally well known. Guidance is available for example from Remington: THE SCIENCE AND PRACTICE OF PHARMACY, 19th Edition, Gennaro (ed.) 1995, Mack Publishing Company, Easton, Pa. Such compositions typically comprise at least one anti-RT aptamer and a pharmaceutically acceptable carrier. The term "pharmaceutically acceptable carrier" refers to any and all coatings, excipients, solvents, dispersion media, absorption delaying agents, and the like, compatible with pharmaceutical administration. Such carriers also include for example sodium chloride, colloidal silica, talc, various polymeric carriers including polyvinyl pyrrolidone, cellulose-based compounds such as carboxymethylcellulose

or methylcellulose, polyvinylpyrrolidone, polyacrylates, and polyethylene glycol. Dosage forms include, for example, oral or sublingual tablets, pellets, micro- and nano-capsules, liposomes, inhalation forms, nasal sprays, and sustained-release preparations. Solutions or suspensions used for administering nucleic acids of the present invention can include one or more of the following components: a sterile diluent such as water for injection, saline solution; fixed oils, polyethylene glycols, glycerine, propylene glycol or other synthetic solvents; antibacterial agents such as benzyl alcohol or methyl parabens; antioxidants such as ascorbic acid or sodium bisulfite; chelating agents such as EDTA; buffers such as acetates, citrates or phosphates and agents for the adjustment of tonicity such as sodium chloride or dextrose. In some embodiments, a pharmaceutical composition can be delivered via slow release formulation or matrix comprising nucleic acids of the present invention or DNA constructs suitable for expression of nucleic acids of the present invention in or around a site within the body.

The invention will be further illustrated by the following figures and examples. However, these examples and figures should not be interpreted in any way as limiting the scope of the present invention.

FIGURES:

Figure 1. A. Schematic representation of a tetrad composed of 4 guanine nucleosides. B. The stacking of 3 tetrads results in the formation of a G-quadruplex structure. C. Typical UV-melting profiles of G4 structure with the T_m defining the specific mid-point transition.

Figure 2: Search of G4 prone sequences in EBOV and MBGV genomes. Bioinformatic search of G4 forming sequences. This graphical representation shows the average score in function of the aligned genomic sequences of from EBOV and MBGV.

EXAMPLE:

Material and Methods:

Bioinformatic analysis: 13 EBOV and 20 MBGV complete genomic sequences were extracted from the Viral Bioinformatics Resource Center. The Fasta files genomes were

aligned using ClustalW program. To detect conserved G4 forming sequences in the genomes alignments we used the algorithm "G4-hunter" that we developed in the laboratory (Bedrat et al, in preparation). It searches for G/C skewness and the presence of G/C blocks in the alignment. It analyses the genome using a sliding window of 25 nucleotides and attribute a score to the first nucleotide of the window. The analysis of sequence conservation was performed using the WebLogo software to generate the LOGO representation.

Preparation of the oligonucleotides: Oligonucleotides were purchased from Eurogentec (Seraing, Belgium) with "Reverse-Phase Cartridge Gold purification". All oligonucleotides were dissolved in 100µM bidistilled water and stored at -20°C. Concentrations were determined by ultraviolet (UV) absorption using the extinction coefficients provided by the manufacturer.

For the UV experiments 4µM oligonucleotides are diluted into 10mM lithium cacodylate buffer and 100 mM KCl

For nuclear magnetic resonance experiments (NMR) the concentration of each samples was typically 100µM in 20 mM potassium phosphate buffer pH 7 containing 70 mM KCl and 10% D2O.

UV-Melting experiments: UV-melting measurements were performed on a Uvikon XS (Secomam) UV-visible spectrophotometer coupled to a water bath temperature-control accessory. A temperature-increase rate of 0.2°C/min was applied and the absorbance values were measured every 1°C. The temperature was measured with an inert glass sensor immersed into a control quartz cell filled with water. The absorbance was monitored at 240 and 295 nm using quartz cells of 0.2 or 1 cm pathlength and 580 µl of volume. Typical UV-melting profiles of G4 structures are represented in figure 1C.

Results

We previously showed that G-quadruplexes oligonucleotides can act as decoys and thus are suitable for the settlement of new anti-viral strategies. This strategy was developed in the context of HIV-1 virus and several G4 sequences (EP14305763). In this strategy we showed that synthetic G4 forming oligonucleotides, derived from the HIV genome, are able to strongly inhibit HIV-1 infectivity in a viral infectivity test realized in vivo with real HIV viruses infecting HeLap4 cells. These G4s might therefore act as decoys and trap crucial

proteins involved in the recognition of the same sequences present in the viral genome. We hypothesize that the presence of G4 sequences in any viral genome might reveal a potential role of these structures in the replication cycles of the virus. If this is the case, the decoy strategy we developed against HIV should also apply for the desired virus.

Using the G4-hunter algorithm we analyzed two alignments of complete genomic sequences from 13 EBOV and 20 MBGV isolates respectively. The average of the 13 or 20 scores obtained for each 25 nt window is depicted in a graphical representation (Figure 2). We consider that the scores higher than 1 in absolute value (-1 or +1) are potentially able to form DNA or RNA G4 structures in the (-) strand (for positive values) or in the (+) strand for the negative values. Six conserved sequences from EBOV and 20 from MBGV were detected as G4 prone sequences. Thorough biophysical analysis by UV-melting experiments revealed that only 5 EBOV and 9 MBGV sequences actually formed thermodynamically stable G4s *in vitro* (Table 1)

Conclusions

We identified 5 EBOV and 9 MBGV G4 forming sequences. As observed for HIV virus, we hypothesize that these sequences might be recognized by viral or cellular proteins involved in important steps of EBOV and MBGV replication cycles. Therefore, as developed in the context of HIV, these oligonucleotides can be used as decoys to inhibit the viral replication of EBOV and MBGV. More generally other G4 forming sequences could also have similar anti-viral properties, in particular the AS1411 G4 forming aptamer (5'GGTGGTGGTGGTTGTGGTGGTGGTGG3') (SEQ ID NO:1) for which we showed recently some inhibitory effects on HIV replication.

Table 1: 5 EBOV and 9 MBGV sequences actually formed thermodynamically stable G4s *in vitro*

EBOV	Start/End	Strand	Sequences (5'-3')	Length	Score	T_m
E1	2298-2323	(+)	CGGTGGGGCGACAGTGGGTGTGCGG (SEQ ID NO:2)	25	1.32	40°C
E2	6980-7005	(-)	CGGGGAGTGGGCCTTCTGGAA (SEQ ID NO:3)	21	1.32	30°C
E3	7480-7509	(+)	GTTTTGGGGACTTGTTGTGGTGGCGGGGT (SEQ ID NO :4)	29	1.41	35°C
E4	10646-10678	(-)	AGGGGTGGAAGGTTTATTGGGCTGGTATTG (SEQ ID NO :5)	30	1.23	30°C
E5	13901-13930	(-)	AGGGGTCATATGGGAGGGATTGAAGGA (SEQ ID NO :6)	27	1.41	20°C

MBGV	Start/End		Sequences	Length	Score	T_m
M1	486-524	(-)	AGAGGGGGAGGATTGGGC (SEQ ID NO :7)	18	1.83	nd
M2	3423-3461	(+)	CGCGGGTTGAGGAGGAGGGA (SEQ ID NO :8)	20	1.3	20°C
M4	6642-6679	(+)	CGGATGGGCTGTGGGCAGTGGTAAAGGT (SEQ ID NO :9)	28	1.04	35°C
M5	6833-6870	(+)	GCGTGCTTGGTTGTGGTGAGGGAGTGGGTGGC (SEQ ID NO :10)	32	1.03	35°C
M8	7190-7242	(+)	TGGGGGTGGGGGAGGGACTGGTGGA (SEQ ID NO :11)	25	2.24	55°c
M8-1	7194-7242	(+)	CAAGATGTTGTGCAGTCGAGTTGGGGGTGGGGGAGGGACTGGTGGAATAC (SEQ ID NO :12)	50	1.18	50°C
M9	7848-7898	(-)	AGAGGGGACTGGTTGGGGTCTG	38	1.47	30°C

			GGTGGTAAATGGTGGA (SEQ ID NO :13)			
M18-1	17341-17375	(-)	TGGCTGAAGGGGAAGGAAGTGGTGCTCGGT (SEQ ID NO :14)	30	1.07	30°C
M19	17346-17375	(-)	TGAAGGGGAAGGAAGTGGTGCTCGGT (SEQ ID NO :15)	26	1.12	20°C

REFERENCES:

Throughout this application, various references describe the state of the art to which this invention pertains. The disclosures of these references are hereby incorporated by reference into the present disclosure.

CLAIMS:

1. A method of treating filovirus infection in a subject in need thereof comprising administering the subject with a therapeutically effective amount of at least one oligonucleotide comprising the sequence as set forth in SEQ ID NO:1 to SEQ ID NO:15.

2. The method of claim 1 wherein the filovirus infection is an Ebola virus infection.

3. The method of claim 2 wherei said Ebola virus is Ivory Coast Ebola virus (ICEBOV), Zaire Ebola virus (ZEBOV or EBOV), Sudan Ebola Virus (SEBOV), or a new strain or species of Ebola virus.

4. The method of claim1 for the treatment of filovirus diseases, in particular Ebola virus disease.

5. An oligonucleotide comprising the sequence as set forth in SEQ ID NO:2 to SEQ ID NO:15.

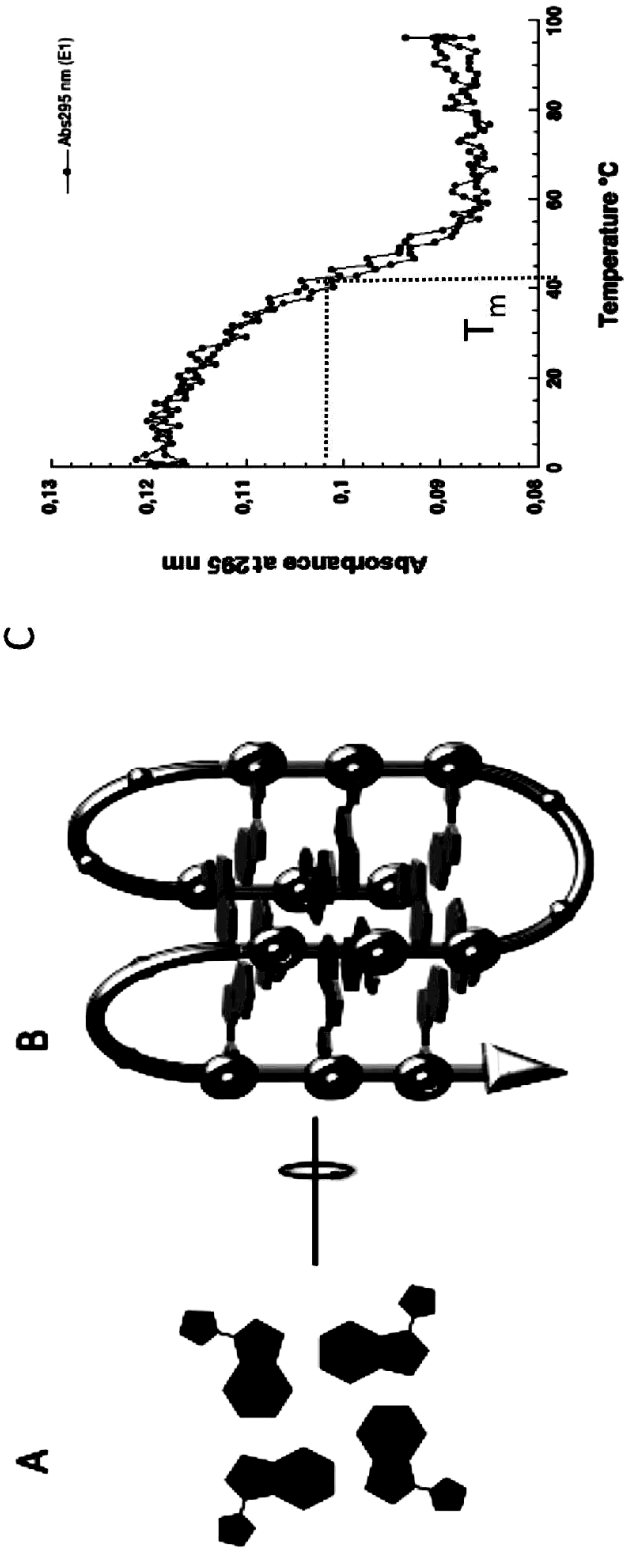

A
B
C
Absorbance at 295 nm
0,13
0,12
0,11
0,1
0,09
0,08
0
20
40
60
80
100
Temperature °C
T_m
Abs295 nm (E1)

Figure 1

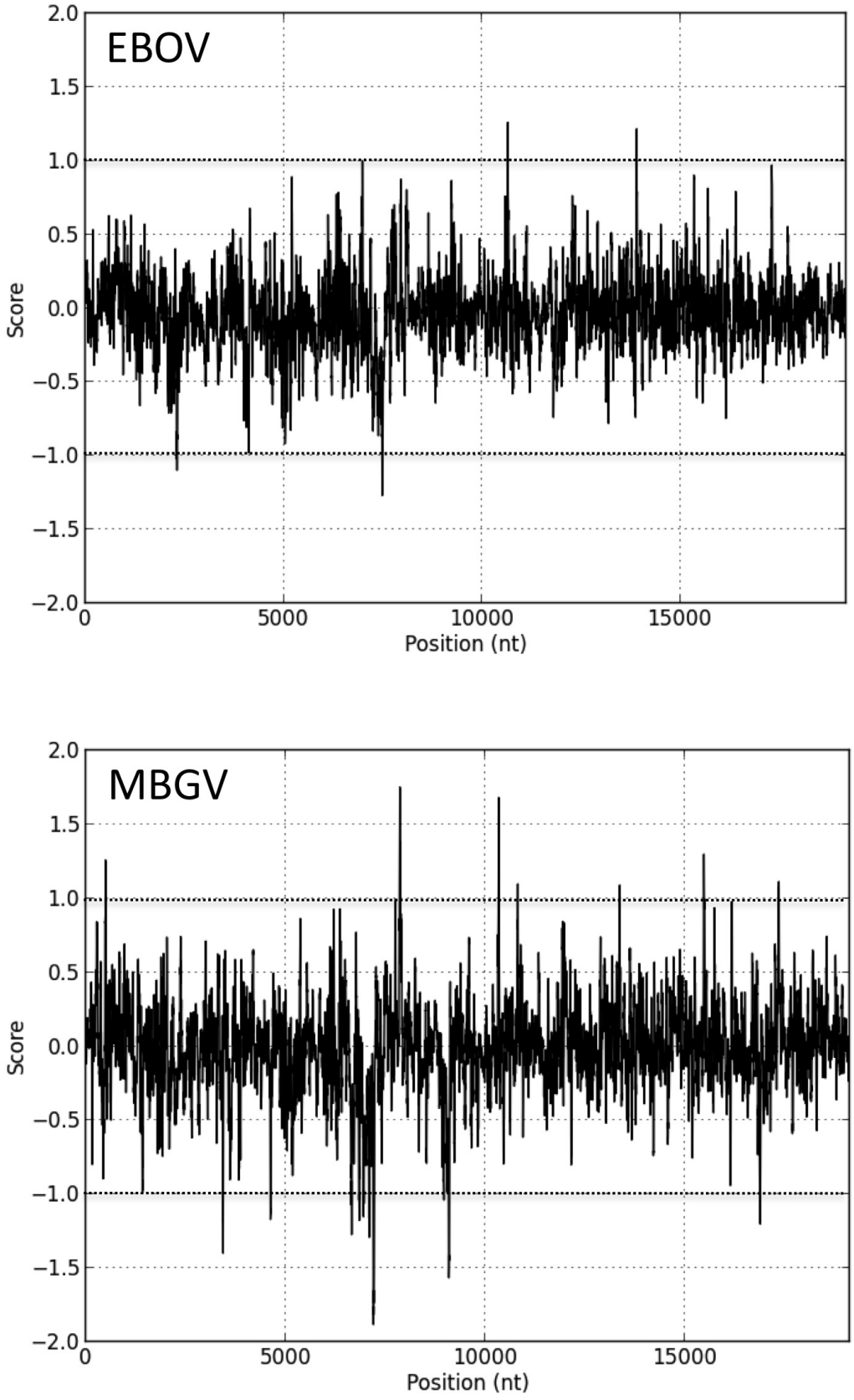

EBOV
Score
Position (nt)
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
0
5000
10000
15000
MBGV
Score
Position (nt)
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
0
5000
10000
15000

Figure 2

Conclusion

Conclusion

This thesis focused on the reevaluation of quadruplex propensity using a new algorithm developed in house. Four-stranded DNA structures are promising drug targets as they appear to in be involved in a number of key biological processes. The increasing interest for G4 makes accurate predictive tools a necessity. Here we developed a completely different algorithm, G4-Hunter, that takes into account G-richness and G-skewness of a considered sequence and gives a score as indicator of G4 propensity. We developed the algorithm using Python programming language. As an input the user can use a FASTA file containing one or more FASTA sequences to be analyzed; the output file containing the PQS sequences and their relative score will be generated at the end of the analysis.

We decided to benchmark this model by using a large number of known sequences. We first selected and reassemble all possible "G4" and "not G4" forming sequences from the literature. This data set was enriched by our unpublished data to obtain 392 sequences. These sequences were separated into two groups "G4" and "not G4" sequences and the score using G4-Hunter was calculated for each sequence. The wilcoxon rank-sum/Mann whitney U-test (null hypothesis: distribution are not different) significantly revealed difference in the distribution of the G4Hscore of the two groups, meaning that the algorithm can distinguish G4 and not G4 sequences by the score calculation method used ($p<2\ 10^{-16}$). The discriminative performance of G4-Hunter and the estimation of the threshold was tested by the Receiver Operating Characteristic (ROC) analysis. The results we obtained were very encouraging. A threshold value of 1 allowed to collect more G4 forming sequences with a rate of false positive no higher than 10%. We concluded at this stage that a threshold of 1 seemed a good choice to search G4 sequences and provided both highest sensitivity and the highest sensibility. Additionally, all score values above 1 performed better than Quadparser.

It should be stated that the dataset used was heavily biased toward G4FS (76%). For this reason, we chose to benchmark G4-Hunter against a different set of sequences, using natural sequences found in the human genome. The human mitochondrial genome was chosen as it is small enough (16.6 kb) to be studied. This genome was previously analyzed to correlate clinically relevant mutations and deletions with non-B DNA propensity. One unique property of this genome is its GC skewness, making it a good candidate to test the performance of G4-Hunter. From the analysis obtained on the data set we chose a threshold of 1 and a window size of 25 nt (a size close to the mean of the dataset sequences size) in order to analyze the human mitochondrial genome.

The results obtained with G4-Hunter were consistent with our expectations: (i) GC skewness was reflected by an average G4Hscore of -0.4 (a long enough random sequence would get a score close to 0) (ii) 1696 new candidates were found, far more than what Quadparser would give. We developed a combination of biophysical methods to accurately assess quadruplex formation *in vitro* for these 165 sequences. Our dataset was enriched by adding 26 sequences identified by quadparser and 7 sequences with a score of 0 to 0.4 as negative controls. The experimental results for the tested sequence (close to 200 different oligonucleotides) are presented in supplementary information ("Supplementary pdf"). We obtained two classes of sequences: G4 folding sequences (128) and not G4 folding sequences (70). The ROC analysis on this mitochondrial dataset shows a discriminating score of 1.2 for a window size of 25 nt. The more the window size increased, the less the number of PQS we found. On the other hand, increasing window size and threshold increased the precision of the detection, but to the cost of a reduction in the number of sequences identified. G4-Hunter is then a good in estimator of the ability of a sequence to fold into a quadruplex.

This algorithm was used here to identify structures expected to adopt an intramolecular fold, but the search parameters do not formally exclude the possibility of quadruplexes of higher molecularity. The estimation of G4FS is characterized by a score that can be handled and adjusted by changing window and threshold values in order to optimize the search. From these conclusions we should be able to enlarge our analysis and search for new PQS in other genomes. We was particularly interested in human pathogens.

The next part of the thesis was therefore focused on G4 sequences in virus. We first evaluated the distribution of G4 motifs through the HIV-1 genome. The idea was to screen the HIV-1 genome using G4-Hunter to identify new conserved PQS. To achieve this goal, modifications are introduced in G4-Hunter source code in order to take into account aligned FASTA sequences. The score calculation has also been modified.
We identified 11 conserved PQS and we checked the folding of these sequences *in vitro*. We demonstrated a highly conserved ability to form G4 in the HIV-1 promoter (PRO-1) implicated in the transcription and regulation of the virus. Structural investigations concluded the stable two-G tetrad antiparallel topology of this PRO-1 G-quadruplex. The nef coding region, fundamental for viral replication and pathogenesis also presents a conserved G-quadruplex forming sequences. The central region of the integrase gene, C-ppt, which is involved in the initiation of the reverse transcriptase present also a highly conserved G4 sequences and it is supposed that this G4 topology helps to maintain in proximity the central region of the RNA genomes during RNA dimerization. The presence of another G4 sequence "gag" came to strengthen the hypothesis that G4 have a global effect on recombination. *In vivo* analysis demonstrated the inhibitory effects of both C-ppt and PRO1 sequences suggesting that they could act as decoys by diverting viral or cellular proteins from their natural targets in the viral genome.

HIV inhibition by G-quadruplex ligands was also observed with high specificity and at low concentration. These observations suggest that G4-based strategies could disturb the HIV-1 replication cycle and pave the way for new anti-HIV therapeutic applications. Our work also allowed the identification of three new G4 sequences in both env and vpr genes. The folding of those sequences into G4 structures could affect the polymerase processing and influence the replication and transcription.

Additionally, the formation of G4 structure could allow different conformations that alters the interaction in the case of vpr.

The conservation of G4 motifs is also found in other viruses. The Ebola and Marburg genomes also contain conserved G4-prone sequences, which we experimentally validated; the majority fold into G4 structure. *In vivo* tests should be of great interest and could help by bringing a new candidate to help to struggle these pathogens. The key of this thesis is G4-Hunter. The easy implementation of this algorithm and the significance of the results obtained on the reference dataset, the human mitochondrial, HIV-1, Ebola and Marburg genomes make it a promising tool.

However, statistical analyses demonstrate some imperfections: there is still room for improvement! The initial values chosen for Guanines blocks could be considered with another scoring strategy, considering floats instead of integers, or be more severe when attributing the bases score. Furthermore, the values assigned for the bases A, T, C in the current version are context independent and obviously an approximation:
for example a GGGCGGGN...NGGGCGGG can form a very stable quadruplex, and the two 1-nt long cytosine loops are actually more favorable than single thymines or adenines: giving a negative score for C and null for A and T is not justified for this family of sequences. Our work enlarge the number of the experimentally validated database sequences, this could be help to simulate the investigation of further improvements.

Annexes

List of Co-signed publication & patents

Publicatons

- Renaud de la Faverie Amandine, Guédin Aurore, **Bedrat Amina,** Yatsunyk Liliya.A. and Mergny Jean-Louis. (2014) Thioflavin T as a fluorescence light-up probe for G4 formation. Nucleic Acids Res, 42, e65.
- Amrane Samir, Kerkour Abdelaziz, **Bedrat Amina,** Vialet Brune, Andreola Marie-Line, and Mergny Jean-Louis. 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., 136(14):5249– 5252, Apr 2014.
- **Bedrat Amina,** Lacroix Laurent and Mergny Jean-Louis (2015) Reevaluation of quadruplex propensity with G4Hunter (soon) .
- **Bedrat Amina,** Amrane Samir, Guédin Aurore and Mergny Jean-Louis (2015) G4s in Viruses, is there a hidden link? (in preparation)

Patents

- S. Amrane, ML Andreola, **A. Bedrat,** JL Mergny Nucleic acids acting as decoys for the treatment of lentivirus infection. EP14305763.6, depot le 23 Mai 2014. (10%)
- S. Amrane, ML Andreola, **A. Bedrat,** JL Mergny Methods and pharmaceutical compositions for the treatment of HIV infection. EP 15 305 329.3, depot le 24/03/15. (10%)
- **A. Bedrat,** S. Amrane, JL Mergny. Methods and pharmaceutical compositions for the treatment of filovirus infections. EP 15 305 367.3, depot le 24/023/15. (33%)

Annexe I

Table 5.5: G-rich sequences extracted from the literature and their G4-Hunter score

Reference	Name	Sequence(5'-3')	Length	Score	Role
[79]	V-545195	GGGGAAAAAGGGGGGGGGGGGGGGGGGGG	31	3.35	-
Unknown	32cc9	CGGGGGGGGGGGGGGGGGCGGGGGAGGGAGGC	32	3.06	-
[79]	IV-1242540	GGGGAGGGGAAGGGGAGGG	19	3.0	S.ceviseae
[222]	Tet22	GGGGTTGGGGTTGGGGTTGGGG	22	2.91	Tetrahymena
[223]	(TG5T)4	TGGGGGTTGGGGGTTGGGGGTTGGGGGT	28	2.86	Tetramolecular
[?]	B	GGGGGATGCGGGGG	14	2.86	-
Unknown	30cc3	CGGAGGGGTGGGGGAGGGGTGGGGGAGGGT	30	2.8	-
Unknown	23cc4	TGGGGGCGGGGAGGGAAGGGGGT	23	2.78	-
[?]	G4CT-pallidum	GGGGCTGGGGCTGGGGCTGGGG	22	2.77	-
Unknown	29cc5	TGGGGTGGGGGTGGGGGTGGTGGTGGGGC	29	2.72	-
-	(TG4T)4	TGGGGTTGGGGTTGGGGTTGGGGT	24	2.67	-
[101]	B-raf	GGGCGGGGAGGGGGAAGGGA	20	2.65	-
Unknown	32cc7	TGGGGTGGGGGTGGAAAGGGGGTGGGGGTGGT	32	2.63	-
[224]	Pu27	TGGGGAGGGTGGGGAGGGTGGGGAAGG	27	2.59	-
[225]	CEB1	AGGGGGGAGGGAGGGTGG	18	2.56	Minisatellites
[226]	VEGF	GGGAGGGTTGGGGTGGG	17	2.53	-
[49]	14h1	TGTGGGGTAGGGGGAGGGGGGAGGGATA	28	2.5	-
[?]	A	GGATGGGGTGGGGAGG	16	2.5	-
199LL-4	26gtel4	AGGGGTTAGGGGTTAGGGGTTAGGGG	26	2.46	-
[49]	20h	TGTGGGGTCGGGGGAGGGGGGAGGGATA	28	2.46	-
[108]	GTERT-110	AGGGGTCGGGTCGGGGCGGGGT	22	2.45	-
Unknown	41cc1	AGGGATGGAACCGGGGAGGGCGGCGGGGGGGGGGGGGGGGGT	41	2.44	-
374LL	CDC25G4-2	CGGGGTGGGCGGGGGGCGAGGGA	23	2.43	-
[108]	GTERT-060	AGGGGAGGGGCTGGGAGGGC	20	2.4	-
[227]	93del	GGGGTGGGAGGAGGGT	16	2.38	Aptamer (HIV-1)
Unknown	35cc8	TGGGGGGAGGGGCGGAGGGGTGGGGTCGCGCGGGT	35	2.37	-
Unknown	29cc2	TGGGGGATGGGGTTGGAATGGGGGCGGGA	29	2.34	-
[226]	Insuline2a	ACAGGGGTGTGGGGACAGGGGTGTGGGG	28	2.29	Promoter
[226]	Oxy3,5/Oxy28	GGGGTTTTGGGGTTTTGGGGTTTTGGGG	28	2.29	Oxytrichanova telomere
[228]	T30695(orT30923)	GGGTGGGTGGGTGGGT	16	2.25	Aptamer (HIV-1)
[187]	Nef8624	GGGGGGACTGGAAGGG	16	2.25	-
[?]	d(G4C2)4	GGGGCCGGGGCCGGGGCCGGGGC	23	2.22	-
Unknown	52cc6	TGGGCGGGGGAGGGAGGGAGGGAGGGGGTTGAGGGTGGGGATGCCAAGGGGT	52	2.19	-
[108]	GTERT-105	CGGGAGGGGTCGGGACGGGGC	21	2.19	-
[49]	UTX	GCCGGGCGGGGAGGGGGGGTCA	22	2.18	-
[229]	HPV32	GGGAGTATGGGTAACGGGGGGGG	23	2.17	-
[230]	BCL-2promoter	GGGCGCGGGAGGAAGGGGGCGGG	23	2.13	P/ANTIP
[223]	A3A	GGGAGGGTTTGGGAGGG	17	2.12	-
[223]	A3T	GGGAGGGTTTGGGTGGG	17	2.12	-
[223]	T3A	GGGTGGGTTTGGGAGGG	17	2.12	-
[223]	T3T	GGGTGGGTTTGGGTGGG	17	2.12	-
[226]	C-Kit2GG	GGGCGGGCGCGAGGGAGGGG	20	2.1	-
[223]	A3C	GGGAGGGTTTGGGCGGG	17	2.06	-
[223]	C3A	GGGCGGGTTTGGGAGGG	17	2.06	-

Reference	Name	Sequence(5'-3')	Length	Score	Role
[223]	C3T	GGGCGGGTTTGGGTGGG	17	2.06	-
[223]	T3C	GGGTGGGTTTGGGCGGG	17	2.06	-
[49]	PSMA2	GCTGGGTTGGGGCGGGGGGAGCGGGACG	28	2.04	-
355LL	CanG3	TGGGGCTGGGGTTACGGGGCCAGTGGGGT	29	2.03	-
[223]	C3C	GGGCGGGTTTGGGCGGG	17	2.0	-
[222]	Gia18	GGGTAGGGTAGGGTAGGG	18	2.0	Giardia
[227]	T95-2T	TTGGGTGGGTGGGTGGGT	18	2.0	Aptamer (HIV-1)
[229]	HPV42	GGGACTATGGGTAACGGGGGGG	22	2.0	-
[?]	J19-TT	TTGGGTGGGTGGGTGGGT	18	2.0	-
Gomez,NAR4	G4TERT2	GGGGGCCTTGGGGCTCGGCAGGGGTGAAAGGGG	33	2.0	-
[223]	12668310-PC12-9	AGTGGGGGTAGGGGATAGGGTAGGC	25	1.96	-
[223]	19apt	GGGTTGGGTGTGGGTTGGG	19	1.95	-
[223]	12668310-PC12-10	GCTGGGGTGTTGGGTGTGGGGGTGA	25	1.92	-
[143]	CSBIIWT	GAAGCGGGGGAGGGGGGGUUUGGUGGAAAU	30	1.9	-
277LL	G4T5	GGGGTTTTTGGGGTTTTTGGGGTTTTTGGG	30	1.9	-
[223]	19G2	GGGTTGGGTTAGGGTTGGG	19	1.89	-
[223]	19T313	GGGTTTGGGTGGGTTTGGG	19	1.89	-
[222]	Gla26	GGGGTCTGGGTGCTGTGGGGTCTGGG	26	1.88	C.glabrata
[228]	T30177(orI100-15)	GTGGTGGGTGGGTGGGT	17	1.88	-
[231]	c-kit2	CGGGCGGGCGCGAGGGAGGGGT	22	1.86	-
[226]	c-Kit1	GGGAGGGCGCTGGGAGGAGGG	21	1.86	c-kitoncogene promoter
[222]	Scer21	GGGTGTGGGTGTGGGTGTGGG	21	1.86	S.ceviseae
[108]	GTERT-051	AGGGCGGGGCCGCGGAAAGGAAGGGGAGGGGC	32	1.84	-
[227]	Pu24T	TGAGGGTGGTGAGGGTGGGGAAGG	24	1.83	c-myconcogene promoter
[229]	HPV52-122	GGGCAGGGGACACAGGGTAGGG	22	1.82	-
374LL	CDC25G4-1	AGGGCGGGACTGAGGGCGGGGC	22	1.82	-
[226]	27KRas	GGGCGGTGTGGGAAGAGGGAAGAGGGG	27	1.81	Promoter
[227]	J19	GIGTGGGTGGGTGGGT	16	1.81	Aptamer
[229]	HPV52	GGGTAGGGCAGGGGACACAGGGTAGGG	27	1.81	-
Gomez,NAR4	VNTR6-1	GGGGTAGGTGGGGATCTGTGGGATTGG	27	1.81	-
[223]	A6A	GGGAGGGTTTTTTGGGAGGG	20	1.8	-
[223]	A6T	GGGAGGGTTTTTTGGGTGGG	20	1.8	-
[223]	T6A	GGGTGGGTTTTTTGGGAGGG	20	1.8	-
[223]	T6T	GGGTGGGTTTTTTGGGTGGG	20	1.8	-
[223]	20G1	GGGTTGGGTTTTGGGTTGGG	20	1.8	-
[223]	20T323	GGGTTTGGGTTGGGTTTGGG	20	1.8	-
[226]	VAV1	GGGCAGGGAGGGAACTGGG	19	1.79	-
[108]	GTERT-056	CGGGGCCGCGGAAAGGAAGGGGAGGGGCTGGGA	33	1.79	-
[226]	32B3(K-Ras)	AGGGCGGTGTGGGAAGAGGGAAGAGGGGGAGG	32	1.78	Promoter
[13]	T30177-T	TGTGGTGGGTGGGTGGGT	18	1.78	-
[227]	c-kit1-87-UP	AGGGAGGGCGCTGGGAGGAGGG	22	1.77	-
[223]	12668310-PC12-16	AGAGTGGGGGGGATGTAGGTGGGTT	25	1.76	-
[223]	12668310-PC12-13	GAGGAGGGAGAATAGGGGTGGGTGG	25	1.76	-
[223]	12668310-PC12-7	GGGTGTGGGAGGTGATGGGGTAGGT	25	1.76	-
[223]	21sAGA	GGGTTTGGGAGAGGGTTTGGG	21	1.76	-
[223]	21sAGT	GGGTTTGGGAGTGGGTTTGGG	21	1.76	-
[223]	21sTGA	GGGTTTGGGTGAGGGTTTGGG	21	1.76	-

Reference	Name	Sequence(5'-3')	Length	Score	Role
[223]	21sTGT	GGGTTTGGGTGTGGGTTTGGG	21	1.76	-
[223]	A6C	GGGAGGGTTTTTTGGGCGGG	20	1.75	-
[223]	C6A	GGGCGGGTTTTTTGGGAGGG	20	1.75	-
[223]	C6T	GGGCGGGTTTTTTGGGTGGG	20	1.75	-
[223]	T6C	GGGTGGGTTTTTTGGGCGGG	20	1.75	-
354LL	URA3G2	AGGGACATGGGTGGGAGGGA	20	1.75	-
[229]	HPV52-223	GGGTAGGGCAGGGGACACAGGGT	23	1.74	-
[223]	cellobiose-1	GCGGGGTTGGGCGGGTGGGTTCGCTGGGCAGGGGGCGAGTG	41	1.73	-
277LL	G4T6	GGGGTTTTTTGGGGTTTTTTGGGGTTTTTTGGG	33	1.73	-
[232]	PS2,M	GTGGGTAGGGCGGGTTGG	18	1.72	-
[222]	Hum21	GGGTTAGGGTTAGGGTTAGGG	21	1.71	-
[223]	21sAAA	GGGTTTGGGAAAGGGTTTGGG	21	1.71	-
[223]	21sAAT	GGGTTTGGGAATGGGTTTGGG	21	1.71	-
[223]	21sAGC	GGGTTTGGGAGCGGGTTTGGG	21	1.71	-
[223]	21sATA	GGGTTTGGGATAGGGTTTGGG	21	1.71	-
[223]	21sATT	GGGTTTGGGATTGGGTTTGGG	21	1.71	-
[223]	21sCGA	GGGTTTGGGCGAGGGTTTGGG	21	1.71	-
[223]	21sCGT	GGGTTTGGGCGTGGGTTTGGG	21	1.71	-
[223]	21sTAA	GGGTTTGGGTAAGGGTTTGGG	21	1.71	-
[223]	21sTAT	GGGTTTGGGTATGGGTTTGGG	21	1.71	-
[223]	21sTGC	GGGTTTGGGTGCGGGTTTGGG	21	1.71	-
[223]	21sTTA	GGGTTTGGGTTAGGGTTTGGG	21	1.71	-
[222]	Par21	GGGTTTGGGTTTGGGTTTGGG	21	1.71	Paramecium
[229]	HT	GGGTTAGGGTTAGGGTTAGGG	21	1.71	-
[223]	12954786-TTF1-2	GATACACGGGCGGAGGAGGTGGGGGGGGGTAGGTGGGTAT	40	1.7	-
[223]	A9A	GGGAGGGTGTAAGTGTGGGAGGG	23	1.7	-
[223]	A9T	GGGAGGGTGTAAGTGTGGGTGGG	23	1.7	-
[223]	C6C	GGGCGGGTTTTTTGGGCGGG	20	1.7	-
[223]	T9A	GGGTGGGTGTAAGTGTGGGAGGG	23	1.7	-
[223]	T9T	GGGTGGGTGTAAGTGTGGGTGGG	23	1.7	-
[233]	H-Bi-G4	GGGACGTAGTGGGGGGACGTAGTGGG	26	1.69	-
[223]	12668310-PC12-2	TGAGGGTCTAGGGTGGTGGGGTGGA	25	1.68	-
[226]	c-Myc	TGAGGGTGGGTAGGGTGGGTAA	22	1.68	-
[223]	12668310-PC12-3	TGATGGATGTGGGGATGCGGGGGCG	25	1.68	-
[13]	T30177-TT(orTB-1)	TTGTGGTGGGTGGGTGGGT	19	1.68	-
[223]	21sAAC	GGGTTTGGGAACGGGTTTGGG	21	1.67	-
[223]	21sACA	GGGTTTGGGACAGGGTTTGGG	21	1.67	-
[223]	21sACT	GGGTTTGGGACTGGGTTTGGG	21	1.67	-
[223]	21sATC	GGGTTTGGGATCGGGTTTGGG	21	1.67	-
[223]	21sCAA	GGGTTTGGGCAAGGGTTTGGG	21	1.67	-
[223]	21sCAT	GGGTTTGGGCATGGGTTTGGG	21	1.67	-
[223]	21sCGC	GGGTTTGGGCGCGGGTTTGGG	21	1.67	-
[223]	21sCTA	GGGTTTGGGCTAGGGTTTGGG	21	1.67	-
[223]	21sCTT	GGGTTTGGGCTTGGGTTTGGG	21	1.67	-
[223]	21sTAC	GGGTTTGGGTACGGGTTTGGG	21	1.67	-
[223]	21sTCA	GGGTTTGGGTCAGGGTTTGGG	21	1.67	-
[223]	21sTCT	GGGTTTGGGTCTGGGTTTGGG	21	1.67	-
[223]	21sTTC	GGGTTTGGGTTCGGGTTTGGG	21	1.67	-
[?]	CatG4	TGGGTAGGGCGGGTTGGGAAA	21	1.67	-

Reference	Name	Sequence(5'-3')	Length	Score	Role
[223]	A9C	GGGAGGGTGTAAGTGTGGGCGGG	23	1.65	-
[230]	BCL-2promoter-modified	GGGCGCGGGAGGAATTGGGCGGG	23	1.65	-
[223]	C9A	GGGCGGGTGTAAGTGTGGGAGGG	23	1.65	-
[223]	C9T	GGGCGGGTGTAAGTGTGGGTGGG	23	1.65	-
[223]	T9C	GGGTGGGTGTAAGTGTGGGCGGG	23	1.65	-
[226]	22Ag	AGGGTTAGGGTTAGGGTTAGGG	22	1.64	Telomere
[223]	22gt	GGGTTAGGGTTAGGGTTAGGGT	22	1.64	-
[223]	22T343	GGGTTTGGGTTTTGGGTTTGGG	22	1.64	-
[223]	12668310-PC12-17	GGTTGGATGTAAGGTTGGAGGGGGG	25	1.64	-
355LL	LysG3	CGGGGTCCGAGGGGATTCCTAAGGGGGTTCTGGGGA	36	1.64	-
[226]	35B1(K-Ras)	AGGGCGGTGTGGGAAGAGGGAAGAGGGGGAGGCAG	35	1.63	-
[223]	21704505-FADGDH-1	TCCGGGGGGCTGGGCAGGGGGGTAACTTTC	30	1.63	-
[108]	GTERT-082	AGGGCCCGGAGGGGGCTGGGCCGGGGA	27	1.63	-
[223]	21sCAC	GGGTTTGGGCACGGGTTTGGG	21	1.62	-
[223]	21sCTC	GGGTTTGGGCTCGGGTTTGGG	21	1.62	-
[229]	HPV3	GGGCTTGGGTGGGCGCTTGGG	21	1.62	-
[**?**]	C	GGAGGGTGGATGG	13	1.62	-
[223]	C9C	GGGCGGGTGTAAGTGTGGGCGGG	23	1.61	-
[223]	25c17b	GGGCCTGTTGGGGTTTGGGTTTGGG	25	1.6	-
[226]	45Ag	GGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGG	45	1.6	-
[223]	25c27b	GGGTTTGGGCCTGTTGGGGTTTGGG	25	1.6	-
[223]	25c37b	GGGTTTGGGTTTGGGCCTGTTGGGG	25	1.6	-
759JL-1	12668310-PC12-15	TGGGTAGGTTCGAGGGGTGGGTGTG	25	1.6	-
[108]	GTERT-076	TGGGAGGGCCCGGAGGGGGCTGGGC	25	1.6	-
[108]	GTERT-089	AGGGGGCTGGGCCGGGGACCCGGGA	25	1.6	-
[186]	Cppt1	TTTTAAAAGAAAAGGGGGGATTGGGGGGTACAGTGCAGGGG	41	1.59	-
Gomez,NAR4	G4TERT1	GGGGTGAAAGGGGCCCTGGGCTTGGG	26	1.58	-
[17]	X3ACT	GGGACTGGGACTGGGACTGGG	21	1.57	-
[223]	TSG4	GGGATTGGGATTGGGATTGGGTT	23	1.57	-
[17]	X3CGC	GGGCGCGGGCGCGGGCGCGGG	21	1.57	-
[226]	21CTA	GGGCTAGGGCTAGGGCTAGGG	21	1.57	Telomere
[17]	X3TCA	GGGTCAGGGTCAGGGTCAGGG	21	1.57	-
[223]	f191t	GGGTGGGTTTTTTTTTGGGTGGG	23	1.57	-
[223]	23T353	GGGTTTGGGTTTTTGGGTTTGGG	23	1.57	-
[62]	Ceb25	AGGGTGGGTGTAAGTGTGGGTGGGT	25	1.56	Minisatellites
[223]	27rap	GGGTGTGTGGGTGTGTGGGTGTGTGGG	27	1.56	-
[223]	26gsc	TTTGTGGGTGTGGGTGTGGGTGTGGG	26	1.54	-
[223]	18617415-Tetracyclines-7	GGGCGGGGGTGCTGGGGGTGGAGTGCTGCGTGCTGCGG	38	1.53	-
[229]	HPV9	GGGAGTGGGAGCGGGAACGGGAACGGGACTGGGA	34	1.53	-
[223]	21sACC	GGGTTTGGGACCGGGTTTGGG	21	1.52	-
[223]	21sCCA	GGGTTTGGGCCAGGGTTTGGG	21	1.52	-
[223]	21sCCT	GGGTTTGGGCCTGGGTTTGGG	21	1.52	-
[223]	21sTCC	GGGTTTGGGTCCGGGTTTGGG	21	1.52	-
277LL	SMG4T5	TCACAGGGGTTTTTGGGGTTTTTGGGGTTTTTGGGGACAA	40	1.52	-
[62]	Ceb26	AAGGGTGGGTGTAAGTGTGGGTGGGT	26	1.5	Minisatellites

Reference	Name	Sequence(5'-3')	Length	Score	Role
[223]	TSG24	AGGGATTGGGATTGGGATTGGGTT	24	1.5	-
[223]	37	AGGGCTAGGGCTAGGGCTAGGG	22	1.5	-
765JL-2	38	AGGGTCAGGGTCAGGGTCAGGG	22	1.5	-
[222]	Tom24	GGGTTAAGGGTTAAGGGTTAAGGG	24	1.5	L.esculentum
[223]	24g	GGGTTAGGGTTAGGGTTAGGGTTA	24	1.5	-
[222]	Ara24-1	GGGTTTAGGGTTTAGGGTTTAGGG	24	1.5	Arabidopsis/Plasmodiume
[223]	24T363	GGGTTTGGGTTTTTTGGGTTTGGG	24	1.5	-
-	24TTG	TTGGGTTAGGGTTAGGGTTAGGGA	24	1.5	-
[223]	21531729-CD16acMet-11	GTAGGTGGGGGACTGGGGACGGGTATGGGCACACGGTAT	39	1.49	-
-	25DDX	GGGCGGGAUAGAGAGCGUGGGCGGG	25	1.48	-
[223]	12668310-PC12-23	GGGTGTGAGAGGTTGAGGGGGTTCG	25	1.48	-
[229]	HPV-5823	GGGCAGGGTAGGGCAATTTAGGG	23	1.48	-
355LL	CanG1	TGGGTGGGACTATTGGGACGGGT	23	1.48	-
[187]	Nef8528	GAGGAGGAGGTGGGT	15	1.47	-
[229]	HPV9-2	GTGGGAGCGGGAACGGGAACGGGA	24	1.46	-
355LL	LysG5	CGGGCGGGTGGGTTGGCCGAGGGT	24	1.46	-
[229]	HPV-5729	GGGAAAGGGTACCTCGAGGGGCCGCGGGG	29	1.45	-
[223]	25c17a	GGGTAGCATTGGGTTTGGGTTTGGG	25	1.44	-
[223]	25c27a	GGGTTTGGGTAGCATTGGGTTTGGG	25	1.44	-
[223]	25T373	GGGTTTGGGTTTTTTTGGGTTTGGG	25	1.44	-
[223]	27CB3	TAAGGGTGGGTGTAAGTGTGGGTGGGT	27	1.44	-
759JL-1	25CD3	TGGGTGGGTTTAATTTTGGGTGGGA	25	1.44	-
[49]	12h	GCTGGGCGAGGGGTGGGAGCAGACGGGCTG	30	1.43	-
277LL	SMG4T6	TCACAGGGGTTTTTTGGGGTTTTTTGGGGTTTTTTGGGGACAA	43	1.42	-
[229]	HPV25-2	GGGAGCGGGACTGGGACCGGGA	22	1.41	-
[223]	12668310-PC12-4	TATGGGGGTGGGTCAGGTTTCGGTA	25	1.4	-
[223]	12668310-PC12-1	TGGTTGGGGATAGAGGTGGGTGTTT	25	1.4	-
355LL	LysG4	CGGGCGGGTGGGTTGGCCGAAGGGT	25	1.4	-
[226]	AKT1	GGGCGGGCGGCTCCGGGCGCGGG	23	1.39	Promoter
[229]	HPV9-1	CGGGAACGGGAACGGGACTGGGA	23	1.39	-
[223]	21531729-CD16acMet-3	GAGCGGGGACGAACACATATGGGGAAGTGGCTTGGGGTGG	40	1.38	-
[223]	24c16a	GGGCCTGTTGGGTTTGGGTTTGGG	24	1.38	-
[226]	PlasC24	GGGTTCAGGGTTCAGGGTTCAGGG	24	1.38	Plasmodiumtelomere
[223]	24c26a	GGGTTTGGGCCTGTTGGGTTTGGG	24	1.38	-
[223]	24c36a	GGGTTTGGGTTTGGGCCTGTTGGG	24	1.38	-
[223]	29CB3	TAAGGGTGGGTGTAAGTGTGGGTGGGTGT	29	1.38	-
-	25TAG	TAGGGTTAGGGTTAGGGTTAGGGATT	26	1.38	-
355LL	CanG2	TGGGTGGGACTATTGGGACAGGGC	24	1.38	-
[49]	14h2	GCCGGGGCGGCTCGGGACGGGGCCCGGGGAGCGTGGGTGGGACC	44	1.36	-
Unknown	39cn1	TGGGCAGGGGAGACACTGGGATCTGAGGGTCTGGGT	36	1.36	-
[223]	23c17d	GGGCCTGTCAGGGTGGGCTAGGG	23	1.35	-
[223]	23c37d	GGGCTAGGGTGGGCCTGTCAGGG	23	1.35	-
[223]	23c27d	GGGTGGGCCTGTCAGGGCTAGGG	23	1.35	-
[223]	cellobiose-2	TAGCGGGTGTGGTGGGTGGGGGAGGCATGGTTTTTGGTAA	40	1.35	-
355LL	LysG6	CGGGCTCACGGGTGGGTATGGGC	23	1.35	-
[223]	f393t	GGGTTTGGGTTTTTTTTTGGGTTTGGG	27	1.33	-
[222]	Chla27	GGGTTTTAGGGTTTTAGGGTTTTAGGG	27	1.33	Chlamydomonas
[229]	HPV9-324	CCGGGAGTGGGAGCGGGAACGGGA	24	1.33	-
[189]	PRO1	TGGCCTGGGCGGGACTGGG	19	1.32	HIVpromoteur

Reference	Name	Sequence(5'-3')	Length	Score	Role
[223]	16518777A-Ricin-4	GGGGGAGGACGCGTAGTGGGGGGCCCATGGTTGTGTGG	38	1.32	-
[79]	IX-356348	GGGTACGGTGGGTAATAAGGGAAGGTATCGGG	32	1.31	-
[223]	21sCCC	GGGTTTGGGCCCGGGTTTGGG	21	1.29	-
[229]	HPV25-1	GGGAGCGGGACTGGGACCGGGACCGGGA	28	1.29	-
Unknown	29cn3	AGGGAAGAGAAAGGGGCCAGGGCCTGGGA	29	1.28	-
[223]	25c17d	GGGCCTGTCAGGGTTTGGGTTTGGG	25	1.28	-
[223]	25c27d	GGGTTTGGGCCTGTCAGGGTTTGGG	25	1.28	-
[223]	25c37d	GGGTTTGGGTTTGGGCCTGTCAGGG	25	1.28	-
[223]	12954786-TTF1-3	TGGCTAGTGGGTAAGGGGCGGGAGGGTGACAGGGCGATCC	40	1.27	-
[?]	AGRO100	GGTGGTGGTGGTTGTGGTGGTGGTGG	26	1.27	-
[49]	GUK1	AGCGGGAGAAGACGGGCTGGGAGGGCGC	28	1.25	-
[223]	18617415-Tetracyclines-1	GGGCAGCGGTGGTGTGGCGGGATCTGGGGTTGTGCGGTGT	40	1.25	-
[223]	cellobiose-3	GTCAAGGTGGGTGGGTGGGGTTGGTTGTTGTTTTGA	36	1.25	-
[223]	25c17c	GGGCCTTTCAGGGTTTGGGTTTGGG	25	1.24	-
[223]	f1E1t	GGGTGGGTTTTTTTTTTTTTTTGGGTGGG	29	1.24	-
[223]	25c27c	GGGTTTGGGCCTTTCAGGGTTTGGG	25	1.24	-
[223]	25c37c	GGGTTTGGGTTTGGGCCTTTCAGGG	25	1.24	-
759JL-4	33CB1	TGGGTGTAAGTGTGGGTGGGTGTAATTGTGGGT	33	1.24	-
[229]	HPV255	GGGAGCGGGACTGGGACCGGGACCGGGACCGGG	33	1.24	-
[234]	c-kit*	GGCGAGGAGGGGCGTGGCCGGC	22	1.23	-
[229]	HPV25-3	GGGACTGGGACCGGGACCGGGA	22	1.23	-
[223]	16518777A-Ricin-3	TGGGGGTCGGGATACGGTCAGTGGTGGTGAGTGGTAACGG	40	1.2	-
277LL	G3T6	GGGTTTTTTGGGTTTTTTGGGTTTTTTGGG	30	1.2	-
[49]	MCM2	GAAGGGACACGGAGGGGCGGGCCAGAGGGTCC	32	1.19	-
Unknown	27cn2	TGGGCCAAGGGCAGAGAAGGGCTGGGA	27	1.19	-
[186]	Cppt2	TTTTAAAAGAAGGGGAGGAATAGGGGATATGA	32	1.19	-
355LL	LysG1	TGGGACCATTGAGGGTGGGAAATTGGACAATGGGGA	36	1.19	-
354LL	URA3G1	CGGGGTGGTGGGCCCAGGGATTGTTAGCGGGT	32	1.19	-
[?]	Cc	AGAGGGTAGATGG	13	1.15	-
[223]	18617415-Tetracyclines-2	GGAGGAACGGGTTCCAGTGTGGGGTCTATCGGGGCGTGCG	40	1.13	-
[223]	32T1H1	GGGTGGGTTTTTTTTTTTTTTTTTTGGGTGGG	32	1.13	-
[235]	TBA	GGTTGGTGTGGTTGG	15	1.13	Trombinaptamer
[49]	NFKB2	TCAGGGTGGGGGCCCCGAGGGCTGGGGCCG	30	1.13	-
[223]	12954786-TTF1-5	TTGGGGTGGGAGGGCGGGTTAACAAAGATAGCGCAACAGG	40	1.13	-
[189]	Promoter	GGGACTTTCCGCTGGGGACTTTCCAGGGAGGCGTGGCCTGGGCGGGACTGGGGAGTGG	58	1.12	-
[223]	7542922-HIV-1-RT-3	CAGGCGTTAGGGAAGGGCGTCGAAAGCAGGGTGGG	35	1.11	-
[223]	f3E3t	GGGTTTGGGTTTTTTTTTTTTTTTGGGTTTGGG	33	1.09	-
[79]	XII-r	GGGGTTACCGCAAAATGGATAGATGGACCGGGGCACACCGGGTAGGGGTCCGGAGGG	57	1.07	-
[223]	15G1	GGTTGGTTAGGTTGG	15	1.07	-
[223]	15GT	GGTTGGTTTGGTTGG	15	1.07	-
Unknown	29cn4	TGGGCCGTGTGGGCAGAGGGCACATGGGC	29	1.07	-
[223]	15984861-NeuropeptideY-12	TGTGAAGGGGGGTACATGACGGGGACTGGCCGGACTACAG	40	1.07	-
[?]	Z-G4	TGGTGGTGGTGGTTGTGGTGGTGGTGTT	28	1.07	-
[79]	X-106443	GGGTCCTCCAAGGGGTAAAACTTACATGGGATGGTGGGGTCACATGGG	48	1.06	-
[223]	18617415-Tetracyclines-3	CGGGAGGGCGGGGTGTGGTATGTATTGAGCGTGGTCCGTG	40	1.05	-
[223]	CLN003	GGAGGGAAAAGTTATCAGGCTGGATGGTAGCTCGGTCGGGGTGGGTGGGTTGGCAAGTGT	60	1.05	-
[223]	f1K1t	GGGTGGGTTTTTTTTTTTTTTTTTTTTTGGGTGGG	35	1.03	-

Reference	Name	Sequence(5'-3')	Length	Score	Role
[185]	GAG	AGCGGGGGGAGAAUUAGAUAAAUGGGAAAAAAUUCGGUUAAGGCCAGGGGGAAA	53	1.02	HIV-1
[223]	21531729-CD16acMet-7	ATCACGTGGTGGGCAAATAACCGGTTGGGGTGGGTCGAGG	40	1.0	-
[223]	21531729-CD16acMet-2	GAGTGCGTAATGGTACGATTTGGGAAGTGGCTTGGGGTGG	40	1.0	-
[223]	16G1	GGTTGGTTTTTGGTTGG	16	1.0	-
[223]	16518777B-Ricin-3	GGAGGCGCGATGTAGGTATGTGAGGGCGGCGGTGGGCG	40	0.97	-
[223]	38T1N1	GGGTGGGTTTTTTTTTTTTTTTTTTTTGGGTGGG	38	0.95	-
[222]	Bom17	GGTTAGGTTAGGTTAGG	17	0.94	-
[?]	Ac	AGATGGAGTGGAGAGG	16	0.94	-
[223]	16518777B-Ricin-2	CCGTAGGTTCGGGGCGGGAGTGGTCCGGAAGGTGGCGTGG	39	0.92	-
[223]	f3K3t	GGGTTTGGGTTTTTTTTTTTTTTTTTGGGTTTGGG	39	0.92	-
[223]	16518777A-Ricin-1	CCGTAGGTTCGGGGTCGGAGTGGTCCGGAAGGTGGCGTGG	40	0.9	-
498JL-5	18GT =18gtel2	AGGTTAGGTTAGGTTAGG	18	0.89	-
[223]	20971648-rHuEPO-a-Ma-2	GATTGAAAGGTCTGTTTTTGGGGTTGGTTTGGGTCAATA	39	0.87	-
277LL	SMG3T6	TCACAGGGTTTTTTTGGGTTTTTTTGGGACAA	39	0.85	-
[223]	f1S1t	GGGTGGGTTTTTTTTTTTTTTTTTTTTGGGTGGG	44	0.82	-
[223]	14744035-HIV-1Nucleocapsidprotein-4	TCGAGGGGTGTGCAAGGCGGGTCAACGGGCCTTATTTGGTGCTTAGGTA	49	0.82	-
[223]	15025912-NS5B-18	GGGGTAGGATAGGGTNTGGAAGGAGGTGCCCCGT	34	0.79	-
[223]	f3S3t	GGGTTTGGGTTTTTTTTTTTTTTTTTTTTTTTTTTGGGTTTGGG	48	0.75	-
[49]	14h3	CCCGGGACGGGGGCCGGCGGGCCACGGGCCC	31	0.71	-
[?]	Bc	AGGAGATGCAGGAG	14	0.71	-
[222]	Asc20	GGCTTAGGCTTAGGCTTAG	19	0.53	Ascaristelomere
[187]	Nef8547	GGTCTTAAAGGTACCTGAGGTCTGACTGG	29	0.38	-
[?]	d(C4G2)4	GGCCCCGGCCCCGGCCCCGGCCCC	24	-2.0	-

Table 5.6: Selected oligonucleotides that do not form G-quadruplexes *in vitro* and their score with G4-Hunter.

Reference	Name	Sequence(5'-3')	Length	Score	Role
[143]	CSBII-295–296mut	GAAGCGGGGGAGGGGGGGUUUGGUAAAAAU	30	1.77	-
-	24RY	CGCGGAAAGGAAGGGGAGGGGCTG	24	1.63	DNA duplex
[17]	X3CGC	GGGCGCGGGCGCGGGCGCGGG	21	1.57	NotG4
[143]	CSBII-305–306mut	GAAGCGGGGGAGGGAAGGUUUGGUGGAAAU	30	1.4	-
[143]	CSBII-312–313mut	GAAGCGGAAGAGGGGGGGUUUGGUGGAAAU	30	1.4	-
170LL	23G13	TGGGGCAGGACGGGTGCCCGGGT	23	1.22	-
[13]	TB-1-2-3-4	TTGTGTGTGTGTGTGGGTGGGT	22	1.09	NotG4
-	22Agm2	AGGGTTAGTGTTAGGTTTAGGG	22	1.09	-
[226]	CGG12	CGGCGGCGGCGGCGGCGGCGGCGGCGGCGGCGG	36	1.0	NotG4
[13]	TB-1-4-7-8	TTGTGGTGGTGTGGGTGTGTGT	22	1.0	NotG4
[13]	TB-1-3-5-7	TTGTGGTGTGGTGTGGTGTGGT	22	0.91	NotG4
[13]	TB-2-3-5-8	TTGGTGTGTGGTGGTGTGGTGT	22	0.91	NotG4
[13]	TB-1-2-3-4-5	TTGTGTGTGTGTGTGTGGTGGGT	23	0.87	NotG4
-	22Agm3	ATGGTTAGTGTTAGGTTTAGGG	22	0.86	-
Gomez, NAR 2004	VNTR6-1mu	AGGATAGGTGAAGATCTGTGGGATTGG	27	0.85	-
-	ura3ct2	AGAGACATGGGTGGAAGAGA	20	0.8	-
[13]	TB-1-2-3-4-5-6	TTGTGTGTGTGTGTGTGTGTGGGT	24	0.75	NotG4
Gomez, NAR 2004	G4TERT2mu	GAGAGCCTTGAAGCTCGGCAGGAGTGAAAGGGG	33	0.73	-
LL, Bioche99	G3Y	TCGAGGTCTTTTCTTTTTTCTTTTAGGGGG	30	0.7	-
[222]	Spom20	GGTTACGGTTACGGTTACGG	20	0.65	NotG4
[222]	Cgi26	GGTGTACTGGTGTACTGGTGTACTGG	26	0.62	NotG4
[222]	Spom23	GGTTACAGGTTACAGGTTACAGG	23	0.57	NotG4
337LL	22Agm4	ATGGTTAGTGTTAGGTTTAGTG	22	0.55	-
-	ura3ct1	CGGTGTGGTGGGCCCAGGTATTGTTAGCGGTT	32	0.53	-
-	12954786-TTF1-4	TTATGGGGATGAAAGTGGTTGTTCGGGTTCGCCACTTCCAC	41	0.51	Aptamer
[226]	dx12(a)	GCGTGAGTTCGG	12	0.5	DNA duplex
-	A46gm8	ATGGTTAGTGTTAGGTTTAGTGTTATGGTTAGTGTTAGTG	40	0.5	-
[13]	TB-1-2-3-4-5-6-7-8	TTGTGTGTGTGTGTGTGTGTGTGTG	25	0.48	NotG4
-	19GA	AGAGAGAGAGAGAGAGAGA	19	0.47	NotG4
-	45LG2T	CATTCGTACGTTGTGTTAGGGTTAGGGTTAGTGAGCAAGTAATAC	45	0.47	-
Gomez, NAR 2004	G4TERT1mu	AGGATGAAAGGAGCCCTGAGCTTGGG	26	0.42	-
-	miniTAR	CCAGAUUUGAGCCUGGGAGCUGUCUGG	27	0.3	-
-	40RT	CATTCGTACGTTAGGAGTAAAAGGAGGATGCAAGTAATAC	40	0.3	-
257LL	TAD1	GGCTCTGGTTAGACCAGATCTGAGCCTGGGAGCTCTCTGGCTAACTAGGGCC	52	0.29	-
206LL	21ACJ	CGCTCACAAACAAAGGGAGCG	21	0.29	-
-	psL1	GAGAGAGAGAGAGAGTTTTTATTAAATATA	30	0.27	Parallel duplex
-	psL2	GAGAGAGAGAGAGAGAAAAATAATTTATAT	30	0.27	Parallel duplex
-	RNAI	GGCACCGUUGGUAGCGGUGCC	21	0.24	-
-	40RTT	CATTCGTACGTTAGGAGTAAAAGTAGGATGCAAGTAATAC	40	0.23	-
257LL	TAD2	GTCGCTGGTTAGACCAGATCTGAGCCTGGGAGCTCTCTGGCTAACTAGCGGC	52	0.21	-
213LL	17AAA	TACACAAAAAAAGGAGA	17	0.18	-
212LL	17ACA	TACACAAACAAAGGAGA	17	0.12	-
206LL	17TGJ	CGCTCTTTGTTTGAGCG	17	0.06	-
-	19AT(a)	ACGTCGATTATAGACGAGC	19	0.05	-
-	Dx	TATAGCTATAA—G-TATAGCTATA	26	0.04	-
-	27sc	GTATTACTTGCTTTTAACGTACGAATG	27	0.04	-
-	CAG12	CAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAG	37	0.0	-
[226]	CTG12	CTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTG	37	0.0	-

Reference	Name	Sequence(5'-3')	Length	Score	Role
-	ds10	CGCAATTGCG	11	0.0	-
-	ds12-CG	CGTTTTAAAACG	13	0.0	-
-	ds12-GC	GCTTTTAAAAGC	13	0.0	-
-	ds14	CGCGAAATTTCGCG	15	0.0	-
-	ds15-h	CGCGAAGCATTCGCG	16	0.0	-
[226]	ds26	CAATCGGATCGAATTCGATCCGATTG	27	0.0	DNA duplex
[226]	dslac	GAATTGTGAGCGCTCACAATTC	23	0.0	-
[226]	GCdx	GCGCGCGCTTTTGCGCGCGC	21	0.0	-
[226]	PS1c(a)	TTTTTTTTTTTATTAAAATTTATAA	25	0.0	-
[226]	PS1c(b)	AAAAAAAAAATAATTTTAAATATT	25	0.0	-
[226]	T30	TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT	31	0.0	-
-	AB16	AAATTTAATTAAATTT	17	0.0	sc DNA duplex
-	AE16	TTTTTTTTAAAAAAAA	17	0.0	sc DNA duplex
-	AG16	AATATATATATATATT	17	0.0	sc DNA duplex
-	CTG5	CTGCTGCTGCTGCTG	16	0.0	trinucleotide
-	CTG7	CTGCTGCTGCTGCTGCTGCTG	22	0.0	trinucleotide
109LL	27dx	TAACCCTAACCCTAAGGGTTAGGGTTA	28	0.0	-
206LL	17TTJ	CGCTCTTTTTTTGAGCG	18	0.0	-
375LL	hp1	TATATACGTACTGTGTTTTCACAGTACGTATATA	35	0.0	-
375LL	hp2	TATATACGTACGGTGTTTTCACCGTACGTATATA	35	0.0	-
375LL	hpctrl	TATATATGTACTGTGTTTTCACAGTACATATATA	35	0.0	-
-	DS26	CAATCGGATCGAATTCGATCCGATTG	27	0.0	-
[?]	d(C4G2)4	GGCCCCGGCCCCGGCCCCGGCCCC	24	-2.0	-
-	24RY	CAGCCCCTCCCCTTCCTTTCCGCG	24	-1.63	-
-	CCG9	CCGCCGCCGCCGCCGCCGCCGCCGCCG	27	-1.0	trinucleotide
-	CCG10	CCGCCGCCGCCGCCGCCGCCGCCGCCGCCG	30	-1.0	trinucleotide
-	CCG15	CCGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCG	45	-1.0	trinucleotide
-	CCG20	CCGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCG			
-		CCGCCGCCGCCGCCGCCGCCG	60	-1.0	trinucleotide
-	26PyTC	CTCCTCTCTTCCCTTCTTCTCTCCTC	26	-0.92	TC tripx 3-2
-	22PyTC3	CCTCTCTCTCCTTCCTCTCTCT	22	-0.82	TC tripx 3-1
108LL	21mix	CTCACAACTCACACTCACACCC	22	-0.82	-
LL, Bioche99	G3R	CCCCCTAAAAGAAAAAAGAAAAGACCTCGA	30	-0.7	-
108LL	21CT	TAACCTAACTAACCCTAATAAA	22	-0.64	-
-	JVITS	GAGAGAGAAACCCCTTTCTCTCTCTTTTCTCTCTCTTT	38	-0.53	Triple TC 3-2
[226]	dx12(b)	CCGAACTCACGC	12	-0.5	DNA duplex
[229]	AC9	ACACACACACACACACAC	18	-0.5	-
[229]	18AC18	ACACACACACACACACACACACACACACACACACAC	36	-0.5	-
-	40YA	GTATTACTTGCATCCTCCTTTTACTCCTAACGTACGAATG	40	-0.3	-
-	RNAII	CGGGUCGCUACCAACGACCCG	21	-0.24	-
-	40YAA	GTATTACTTGCATCCTACTTTTACTCCTAACGTACGAATG	40	-0.23	-
-	R06	UCAACACGGUCCCAGACGUGUUGA	24	-0.21	-
LL, Bioche99	G3TC18	TTTTCTTTTTTCTTTTCT	18	-0.17	-
-	T1RY+	CACTCCCTATCAGTGATAGAGAGAGAGAAAAAAAGAGAAGA			
-		TCTGAGCTCGGTACCCT	58	-0.14	-
213LL	17TGT	TCTCCTTTGTTTGTGTA	17	-0.12	-
-	19AT(b)	TGCAGCTAATATCTGCTCG	19	-0.05	-
-	27sg	CATTCGTACGTTAAAAGCAAGTAATAC	27	-0.04	-
-	27SGT	CATTCGTACGTTTTTAGCAAGTAATAC	27	-0.04	-

Table 5.7: G-rich sequences in the human mitochondrial genome predicted by G4-Hunter for a window of 25 nucleotides and score higher than 1

Name	Score	IDS	Tm	Tm	TDS	CD	RMN			Thioflavine	G4/Not G4G4	Type
			Tm				25°C	4°C	38°C			
1	?	No		No	No	No	ND			+	Not G4	
2	-1.0	No		Yes	No	No	Yes			+++	G4	Unstable
3	-2.46	Yes		Inconclusive	Yes	Parallel	Yes			++	G4	
4	-1.1	Yes	34°C	Yes(<37°C)	Yes	Anti-Parallel	ND			++	G4	Unstable
5	-1.57	Yes	57°C	Yes	Yes	Parallel	ND			++	G4	
5a	-1.39	Yes(-)		Yes(-)	Yes(-)	Parallel	ND			+++	G4	
6	-1.0	No		No	No	Parallel	Yes			++	G4	Prbably tetramolecular
7	-1.08	Yes		Yes	Yes	Parallel	ND			++	G4	
8	-1.0	No		No	No	No	ND			+	Not G4	
9	-1.88	Yes	60°C	Yes	Yes	Mixed	ND			++	G4	
10	-1.04	No		No	No	No	ND			+	Not G4	
11	-1.08	No		No	No	No	ND			+	Not G4	
12	-1.22	Yes	37°C	Yes	No	No	Yes			++	G4	
13	-1.4	Yes	50°C/H	Yes	Yes(-)	Parallel	Yes			++	G4	
14	-1.15	No		No	No	No	ND			++	Not G4	
15	-1.11	No		Yes(-)	No	Parallel	No			+	Not G4	
16	-1.0	No		No	No	No	No			++	Not G4	
17	-1.0	No		No	No	No	No			++	Not G4	
18	-1.92	Yes	55°C	Yes	Yes	Mixed	Yes			++	G4	
19	-0.89	No		No	No	No	ND			(+/-)	Not G4	
20	-0.93	No		No	No	Mixed	ND			+	Not G4	
21	-1.52	Yes	45°C	Yes	Yes	Mixed	ND			++	G4	
22	-1.0	No		Yes(<37°C)	No	No	Yes	Yes		++	G4	Unstable
23	-1.12	Yes	37°C	Yes(<37°C)	Yes(-)	Mixed	ND			++	G4	Unstable
24	-1.33	Yes		Yes(<37°C)	Yes	Parallel	Yes			++	G4	Unstable
24a	-0.88	Yes		Yes(<37°C)	Yes (-)	Parallel	Yes			++	G4	Unstable
25	-0.95	Yes		Yes(<37°C)	No	Parallel	Yes			+	G4	Unstable
26	-1.16	Yes(-)	37°C	Yes(<37°C)	No	Parallel	Yes		Yes	++	G4	
27	-1.26	Yes	40°C	Yes	Yes	Mixed	Yes			++	G4	
28	-1.04	Yes	40°C	Yes	No	Mixed	No	Yes		(+/-)	G4	Minor species
29	-2.65	Yes	70°C	Yes	Yes	Parallel	Yes			+++	G4	
30	-1.18	Yes	37°C	Yes	Yes	Antiparallel	ND			++	G4	
31	-0.93	No		No	No	No	Yes/CG-AT			+++	G4	Competition
32	-1.12	No		No	No	No	Yes			+	G4	Special RMN
33	-1.04	Yes(-)	20°C	Yes(<37°C)	Yes	Antiparallel	No	Yes		++	G4	Unstable
34	-1.04	No		No	No	No	ND			+	Not G4	
35	-1.03	No		Yes(-)	No	No	Yes			++	G4	Unstable
36	-1.14	No		Yes(<37°C)	No	Parallel	No	Yes		++	G4	Unstable
38	-1.44	No		No	No	No	ND			++	Not G4	
39	-1.03	Yes		Yes(<37°C)	Yes (-)	No	ND			++	G4	Unstable

Name	Score	IDS		Tm	TDS	CD	RMN 25°C	RMN 4°C	RMN 38°C	Thioflavine	G4/Not G4G4	Type
40	-1.14	No		No	No	No	ND			++	Not G4	
41	-1.11	No		No	No	No	Yes	Yes		+++	G4	
42	-1.21	No		Yes (-)	No	Parallel	No	Yes		+	G4	Concl basée sur RMN 4°C
43	-1.08	Yes	40°C	Yes	Yes	Parallel	ND			++	G4	
44	-1.16	No		No	No	No	ND			(+/-)	Not G4	
45	-1.0	No		No	No	No	??	???		+	Not G4	
46	-1.04	No		No	No	No	No			++	Not G4	
47	-1.0	No		Yes(<37°C)	No	No	Yes	Yes		+	G4	Unstable
48	-1.17	Yes(-)		Yes(<37°C)	Yes(-)	Parallel	ND			++	G4	Unstable
49	-1.23	No		No	No	No	CG-AT			++	Not G4	
50	-1.08	Yes (-)		Yes (-)	No	Mixed	Yes/CG			+	G4 /??	Compet/Unstable
51	-1.0	No		No	No	No	ND			++	Not G4	
52	-1.18	Yes		Yes(<37°C)	No	Parallel	Yes(-)			+++	G4	Unstable
53	-0.9	No		Yes(<37°C)	No	No	No	Yes		++	G4	Unstable
54	-1.0	No		Yes(<37°C)	No	No	No	Yes		++	G4	Unstables
55	-1.73	Yes	60°C	Yes	Yes	Parallel	ND			++	G4	
56	-1.15	No		No	No	No	Yes(-)	Yes		++	G4	Based on NMR
57	-1.1	No		No	No	No	ND			+	Not G4	
58	-1.0	No		No	No	No	ND			+	Not G4	
59	-1.03	Yes (-)	40°C	Yes	Yes (-)	Mixed	Yes	Yes		++	G4	
60	-1.29	No		No	No	Parallel	ND			++	Not G4	
61	-1.2	No		No	No	No	ND			+	Not G4	
62	-1.0	No		Yes(<37°C)	No	Parallel	No	Yes		+++	G4	Unstable
63	-1.0	No		Yes(<37°C)	No	Parallel	No			++	G4	Unstable
64	-1.58	Yes	40°C	Yes	Yes(-)	Mixed	ND			++	G4	
65	-1.0	No		No	No	No	ND			++	Not G4	
66	-1.1	Yes (-)	30°C	Yes(<37°C)	Yes	No	No	Yes		++	G4	Unstable
67	-0.93	No		No	No	No	No			+++	Not G4	
68	-1.04	Yes (-)		Yes(<37°C)	No	No	Yes			++	G4	Unstable
69	-0.96	No		No	No	No	ND			+	Not G4	
70	-1.04	No		No	No	Paralle	No			++	Not G4	
71	-1.52	Yes	40°C	Yes	Yes	Paralle	Yes			+++	G4	
72	-1.02	No	40°C	Yes(?)	No	No	Yes			++	??	
73	-1.0	No		Yes(<37°C)	No	No	No	Yes		++	G4	Unstable
74	-1.13	No		No	No	No	ND			+	Not G4	
75	-1.08	No		No	No	No	No			++	Not G4G4	
76	-1.41	Yes	45°C	Yes	Yes	Paralell	Nd			++	G4	
77	-1.09	No		Yes(-)	No	No	/CG-AT			++	Not G4	(minori)
78	-1.3	Yes	37°C	Yes	Yes	Paralle	Yes			+++	G4	
79	-1.04	No		No	No	No	ND			++	Not G4	
80	-1.24	Yes	30°C	Yes(<37°C)	Yes	Antiparalle	ND			++	G4	Unstable
81	-1.63	Yes (-)		Yes	Yes(-)	Pa	Yes			++	G4	
82	-1.0	No		Yes(<37°C)	No	Parallel	Yes			+++	G4	Unstable

Name	Score	IDS		Tm	TDS	CD	RMN			Thioflavine	G4/Not G4G4	Type
							25°C	4°C	38°C			
83	-1.13	No		Yes(<37°C)	No	Paralle	Faire			++	Not G4	
84	-1.0	No		Yes(<37°C)	Yes	Paralle	Yes			++	G4	Unstable
85	-1.48	No		Yes(<37°C)	Yes (-)	??	Yes			++	G4	Unstable
86	-1.5	No		Yes(<37°C)	No	??	Yes			++	G4	
87	-1.15	No		No	No	Parallel	Yes	Yes		++	G4	
88	-1.45	Yes	40°C	Yes	Yes	Mixed	ND			+++	G4	
89	-1.27	No	40°C	Yes	No	Mixed	Yes			++	G4	
90	-1.52	Yes	50°C	Yes	Yes	parallel	ND			+++	G4	
91	-1.3	Yes (-)	37°C	Yes	No	Mixed	Nd			++	G4	
92	-1.03	Yes (-)		Yes(<37°C)	No	Parallel	Yes			++	G4	Unstable
93	-1.0	No		No	No	No	ND			++	Not G4	
94	-0.9	No		No	No	No	ND			++	Not G4	
95	-1.85	Yes	55°C	Yes	Yes	Parallel	Nd			++	G4	
96	-1.24	No		Yes(<37°C)	No	Parallel	No	Yes		+/-	G4	Unstable
97	-1.37	No		Yes(<37°C)	Yes	No	ND			++	G4	Unstable
98	-1.08	No		No	No	No	ND			+/_	Not G4	
99	-1.14	No		No	No	No	ND			++	Not G4	
100	-1.16	Yes (-)	40°C	Yes	No	No	faire			++	G4	
101	-1.16	No		No	No	No	No			++	Not G4	
102	-1.03	No		Yes(<37°C)	No	Parallel	Yes			++	G4	Unstable
103	-0.96	No		No	No	Parallel	Nd			++	Not G4	
104												
105	-1.31	Yes	50°C	Yes	Yes	Mixed	Yes			++	G4	
106	-1.08	Yes		Yes(<37°C)	Yes	No	ND			++	G4	Unstable
107	-1.0	Yes		No	No	No	No			++	Not G4	
108	-1.32	No		Yes(<37°C)	Yes	No	Yes			++	G4	Unstable
109	-1.06	No		Yes(<37°C)	Yes	No	Yes/CG			+++	G4	Unstable+COMP
110	-1.0	No		No	No	No	CG-AT			+	Not G4	
111	-1.13	No		No	No	No	AT			++	Not G4	Competition
112	-1.9	Yes	65°C	Yes	Yes	Parallel	Nd			+++	G4	
113	-1.0	Yes		Yes(<37°C)	Yes	Mixed	Yes			++	G4	Unstable
114	-1.43	Yes	40°C	Yes	Yes	Yes	Mixed			+++	G4	
115	-1.19	No		No	No	No	ND			++	Not G4	
116	-1.0	No		No	No	No	ND			++	Not G4	
117	-1.21	Yes	40°C	Yes	Yes	Mixed	Yes			++	G4	
118	-1.0	No		No	No	No	No	Yes		+++	G4	
119	-1.04	Yes	39°C	Yes	Yes	Antiparallel	Yes			++	G4	
120	-1.15	Yes (-)		No	No	Paralle	Yes /CG			++	G4	COMPETITION
121	-1.13	No		Yes	No	Mixed	Yes		Yes	++	G4	
122	-1.28	Yes	37°C	Yes	Yes	Mixed	ND			++	G4	
123	-1.29	Yes	45°C	Yes	Yes	Parallel	ND			++	G4	
124*	-1.23	No		No	Yes	Parallel	Yes			+++	G4	
125*	-1.14	No		No	No	Parallel	Yes		Yes	++	G4	
126	-1.0	No		Yes(-)	No	Parallel	Yes			++	G4	Unstable
127	-1.0	Yes(-)	35°C	Yes	No	Parallel	Yes			++	G4	Unstable
128	-0.92	No		No	No	No	ND			++	Not G4	

Name	Score	IDS		Tm	TDS	CD	RMN			Thioflavine	G4/Not G4G4	Type
							25°C	4°C	38°C			
129	-1.17	No		No	No	Mixed	ND			++	Not G4	
130	-1.59	Yes	60°C	Yes	Yes	Mixed	ND			++	G4	
131	-1.0	No		No	No	Parallel	ND			++	Not G4G4	
132	-1.82	Yes	70°C	Yes	Yes	PARALLEL	Nd			++	G4	
133	-0.93	No		No	No	No	No			++	Not G4	
134	-0.96	No		No	No	No	ND			++	Not G4	
135	-1.0	Yes (-)		Yes(<37°C)	No	Mixed	CG			++	G4	COMPETITION+Unstable
136	-1.12	Yes	37°C	Yes	No	Mixed	Yes			++	G4	
137	-1.07	Yes	40°C	Yes	Yes	Antiparallel	ND			++	G4	
138	-1.19	Yes		Yes(<37°C)	Yes (-)	Parallel	Yes			++	G4	Unstable
139	-0.96	No		No	No	No	Yes/CG-AT			++	Not G4	
140	-1.17	Yes		Yes(-)	Yes	Mixed	Yes			++	G4	Unstable
141	-1.38	Yes	37°C	Yes	Yes	Mixed	ND			++	G4	
142	-1.5	Yes	40°C	Yes	Yes	Mixed	ND			++	G4	
143	-1.31	Yes	40°C	Yes	Yes	Parallel	Yes			++	G4	SABLE
144*	-1.08	Yes		Yes(<37°C)	Yes	Antiparallel	ND			++	G4	Unstable
145*	-1.21	Yes (-)		Yes(<37°C)	Yes	Mixed	Yes			+++	G4	Unstable
146	-1.0	No		Yes(?<37°C)	No	No	No			++	Not G4	
147	-3.0											
148	-1.08	No		Yes(<37°C)	No	Mixed	Yes			++	G4	Unstable
149	-1.35	Yes	40°C	Yes	Yes	Mixed	ND			++	G4	
150	-1.0	No		No	No	Parallel	Yes ??			++	Not G4	
151	-1.26	No		Yes(<37°C)	No	Parallel	Yes			+	G4	
152	-1.42	No		No	No	Parallel	Yes			++	G4	
153	-0.96	No		Yes(<37°C)	No	No	Yes			++	Not G4	
154	-1.48	Yes	37°C	Yes	Yes	Mixed	ND			++	G4	
155*	-1.54	Yes		No	No	Parallel	Yes		Yes	++	G4	
156*	-1.0	No		Yes(<37°C)	No	No	Yes	Yes		++	G4	Unstable
157	-1.26	Yes	37°C	Yes	Yes	Parallel	ND			+++	G4	
158	-0.96	Yes		No	No	Parallel	ND			++		
159	-1.13	Yes (-)		Yes(<37°C)	Yes (-)	Mixed	Yes			++	G4	Unstable
160	-1.45	Yes	60°C	Yes	Yes	Mixed	ND			++	G4	
161	-1.12	Yes		Yes(<37°C)	Yes	Mixed	ND			++	G4	Unstable
162	-0.94	No		No	No	No	No			++	Not G4 G4	
163	-3.0											
164	-2.07	Yes	63°C	Yes	Yes	Mixed	ND			++	G4	
165	-1.46	Yes	42°C	Yes	Ues	Antiparallel	ND			++	G4	
166	-1.43	No		No	No	Parallel	Yes			++	G4	
167	-1.26	Yes (-)		Yes (-)	Yes	Mixed	Yes			++	G4 ??	

Table 5.8: Selected oligonucleotides with a potential quadruplex-forming. All the sequences are experimentally validated and their score with G4-Hunter is shown.

Name	Score	IDS	Tm	TDS	CD	RMN	Thioflavine	G4/NoTG4
0.5_1	1.26	Yes	Yes	Yes	Parallel	ND	+++	G4
0.5_2	1.11	Yes	Yes	Yes	AntiParallel	ND	++	G4
0.5_3	0.71	Yes	Yes	Yes	AntiParallel	ND	++	G4
0.5_4	1.0	Yes	Yes	Yes	AntiParallel	ND	++	G4
0.5_5	1.18	Yes	Yes	Yes	Mixte	ND	+++	G4
0.5_6	1.65	Yes	Yes	Yes	Mixte	ND	+++	G4
0.5_7	0.88	Yes	No	No	No	ND	++	Not G4
0.5_8	1.22	No	No	Yes	AntiParallel	ND	++	G4
0.5_9	0.73	No	No	No	Mixte	ND	++	NoT
0.5_10	1.16	Yes	Yes	Yes	Mixte	ND	+++	G4
0.5_11	1.18	Yes	Yes	Yes	Mixte	ND	+++	G4
0.5_12	0.84	Yes	Yes	Yes	Mixte	ND	+++	G4
0.5_13	0.92	Yes	Yes	Yes	Mixte	ND	++	G4
0.5_14	0.95	Yes	Yes	Yes	AntiParallel	ND	++	G4
0.5_15	0.89	Yes	Yes	Yes	AntiParallel	ND	+++	G4
0.5_16	1.41	Yes	Yes	No	Mixte	ND	++	G4???
0.5_17	0.78	Yes	Yes	No	No	ND	++	G4
0.5_18	1.12	Yes	Yes	Yes	Mixte	ND	+++	G4
0.5_19	0.84	No	No	No	No	ND	++	Not G4
0.5_20	0.72	No	No	No	No	ND	++	Not G4
0.5_21	1.21	Yes	Yes	Yes	AntiParallel	ND	++	G4
0.5_22	0.7	Yes	Yes	Yes	AntiParallel	ND	++	G4
0.5_23	0.94	Yes	Yes	Yes	Mixte	RMN	+++	??
0.5_24	0.6	No	No	No	No	ND	++	Not G4
0.5_25	0.96	No	No	No	No	RMN	+++	Not G4
0.5_26	0.63	No	No	No	No	ND	+	Not G4
0.5_27	0.85	Yes	Yes	Yes	AntiParallel	ND	++	G4
0.5_28	0.43	Yes	Yes	Yes	AntiParallel	ND	+++	G4
0.5_29	0.96	Yes	Yes	Yes	AntiParallel	ND	+++	G4
0.5_30	1.03	Yes	Yes	Yes	Mixte	ND	+++	G4
0.5_31	0.61	No	No	No	No	RMN	++	Not G4
0.5_32	0.83	No	No	No	AntiParallel	ND	+	Not G4
0.5_33	0.52	No	No	No	No	ND	+	Not G4
0.5_34	0.8	No	No	No	No	ND	+	Not G4
0.5_35	0.82	Yes	Yes	Yes	AntiParallel	ND	+	G4
0.5_36	0.96	No	No	No	No	ND	++	Not G4
0.5_37	1.11	Yes	Yes	Yes	Parallel	ND	+++	G4
0.5_38	0.0	No	No	No	No	ND	+	Not G4
0.5_39	0.0	No	No	No	No	ND	++	Not G4
0.5_40	0.16	No	No	No	No	ND	++	Not G4
0.5_41	0.16	No	No	No	No	ND	++	Not G4
0.5_42	0.36	No	No	No	No	ND	++	Not G4
0.5_43	0.4	No	No	No	No	ND	++	Not G4
0.5_44	0.4	No	No	No	No	ND	++	Not G4

Table 5.9: Number of hits by kbp of the sequenced genome obtained with the G4-Hunter tool using a window of 25 and the threshold indicated in the first column. Reference genomes are respectively hg19(H.s.), mm10(M.m.), dm3(D.m.), ce10(C.e.), ddAX4(D.d.), SacCer3(S.c.), Ecol.(E.c.), TAIR9(T.a.), rheMac3(Mac), panTro3(Chimp), bosTau6(Cow.), susScr3(Pig), canFam3(Dog), rn5(Rat), galGal4(Chicken), danRer7(Zebrafish) , apiMel2(Bee) and MSU7(rice). Note that the E.c. reference genome contains 13 genomes of different E.Coli strains. Note that to calculate the length of the sequenced genome, unattributed bases N have been excluded

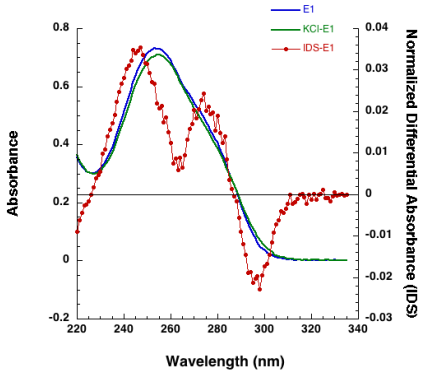
Threshold	H.sapiens	M.musculus	D.melanogaster	C.elegans	D.discoideum	S.cerevisiae	E.Coli	A.thaliana
1	2.425	2.329	1.575	0.817	0.289	0.698	1.301	0.704
1.25	1.010	1.027	0.609	0.268	0.078	0.151	0.285	0.158
1.5	0.502	0.571	0.300	0.112	0.031	0.042	0.075	0.042
1.75	0.247	0.344	0.150	0.052	0.014	0.013	0.019	0.013
2	0.119	0.215	0.076	0.029	0.007	0.005	0.006	0.005

Threshold	Macaque	Chimpanzee	Cow	Pig	Dog	Rat	Chicken	Zebrafish	Bee	Rice
1	2.425	2.391	2.508	2.524	2.791	2.457	2.079	1.137	1.281	2.279
1.25	0.995	0.989	1.132	1.181	1.340	1.083	0.836	0.412	0.639	1.029
1.5	0.478	0.486	0.615	0.651	0.739	0.576	0.395	0.190	0.370	0.516
1.75	0.224	0.238	0.327	0.363	0.412	0.325	0.196	0.091	0.213	0.254
2	0.101	0.114	0.174	0.199	0.228	0.189	0.101	0.047	0.116	0.119

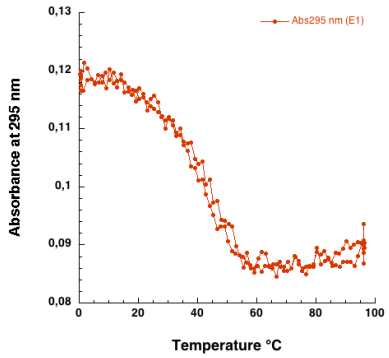
Annexe II

Name: Ebola 1 **Sequence:** $^{5'}$*CGGTGGGGCGACAGTGGGTGTGCGG* $^{3'}$ **Score:** 1.32

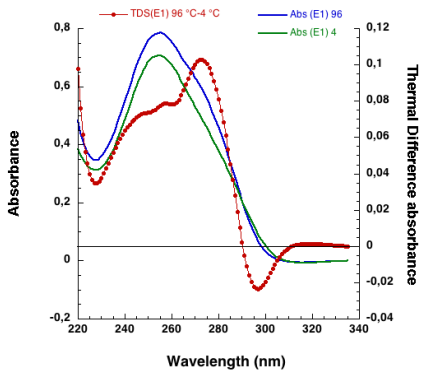
(a)



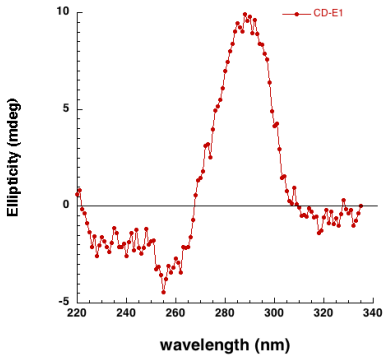
(b)



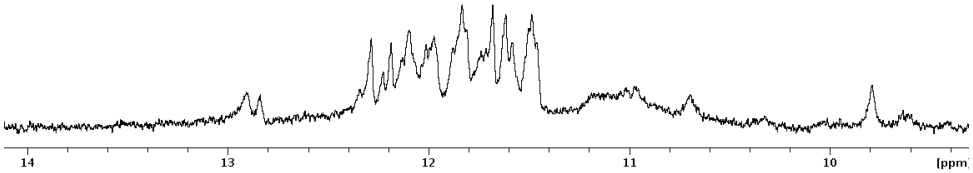
(c)



(d)



(e)



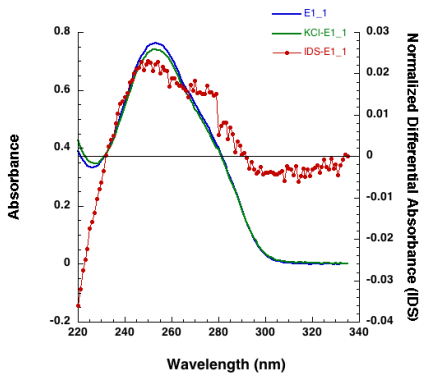
In vitro characterization of the selected candidates: **a)** Normalized Isothermal Difference Spectra (IDS) resulting from the difference between the absorbance recorded at 25°C before and after annealing in 100 mM KCl. **b)** Thermal melting transition (Tm) profiles measured at 240 and/or 295 nm (4 µM strand concentration, 10mM sodium cacodylate (pH= 7.0) and 100 mM KCl). **c)** Normalize differential spectra (TDS) in the 220 and 335 nm region (4 µM strand concentration, 10mM lithium cacodylate (pH= 7.0) and 100 mM KCl). **d)** Circular dichroism spectra (CD) (4 µM strand concentration, 10mM sodium cacodylate (pH= 7.0) and 100 mM KCl, 20°C). **e)** 1D Imino proton spectra in a 20 mM potassium phosphate buffer pH 6.9 with 70 mM KCl at 25°C. The presence of peacks between 10 and 12 ppm suggests G4 formation.

Table 1: Results interpretation of Ebola 1

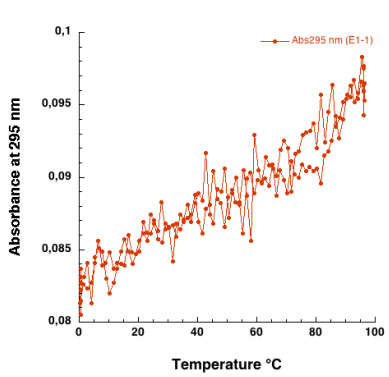
Technique	IDS	Tm	TDS	CD	NMR	**Conclusion**
Result (G4 ?)	Yes	Yes	Yes	No	Yes	**G4 (Stable)**

Name: Ebola 1-1 **Sequence:** $^{5'}CGGGAGCCGGTGGGGCGACAGTGGG^{3'}$ **Score:** 1.36

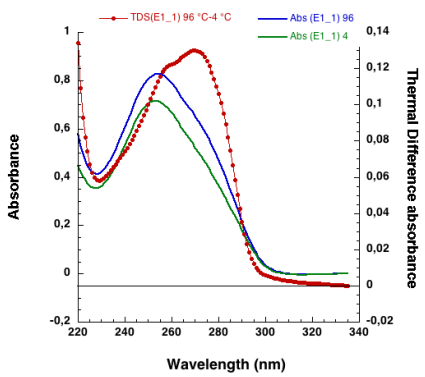
(a)



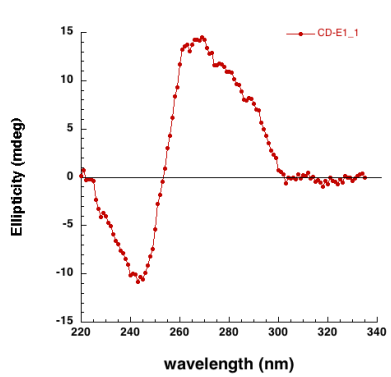
(b)



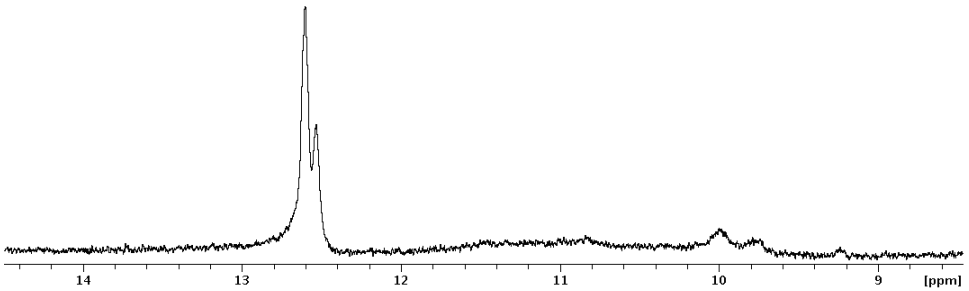
(c)



(d)



(e)



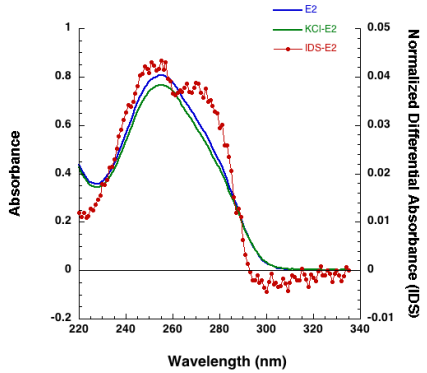
In vitro characterization of the selected candidates: **a)** Normalized Isothermal Difference Spectra (IDS) resulting from the difference between the absorbance recorded at 25°C before and after annealing in 100 mM KCl. **b)** Thermal melting transition (Tm) profiles measured at 240 and/or 295 nm (4 µM strand concentration, 10mM sodium cacodylate (pH= 7.0) and 100 mM KCl). **c)** Normalize differential spectra (TDS) in the 220 and 335 nm region (4 µM strand concentration, 10mM lithium cacodylate (pH= 7.0) and 100 mM KCl). **d)** Circular dichroism spectra (CD) (4 µM strand concentration, 10mM sodium cacodylate (pH= 7.0) and 100 mM KCl, 20°C). **e)** 1D Imino proton spectra in a 20 mM potassium phosphate buffer pH 6.9 with 70 mM KCl at 25°C. The presence of peacks between 10 and 12 ppm suggests G4 formation.

Table 2: Results interpretation of Ebola 1-1

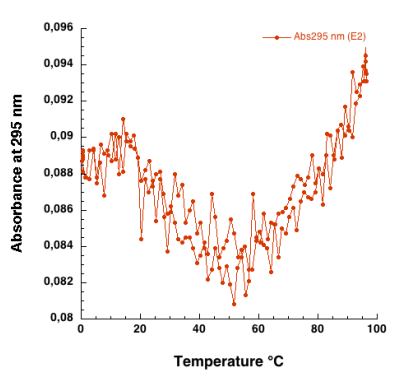
Technique	IDS	Tm	TDS	CD	NMR	**Conclusion**
Result (G4 ?)	No	No	No	Mixed	No	**Not G4**

Name: Ebola 2 **Sequence:** $^{5'}$*CGGGGAGTGGGCCTTCTGGGAA* $^{3'}$ **Score:** 1.32

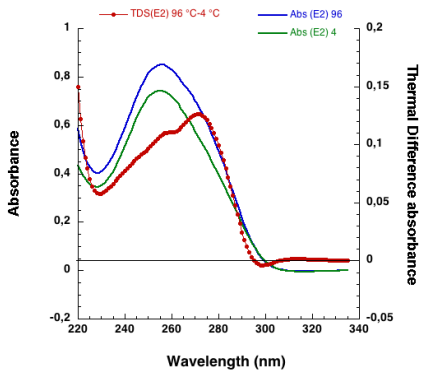
(a)



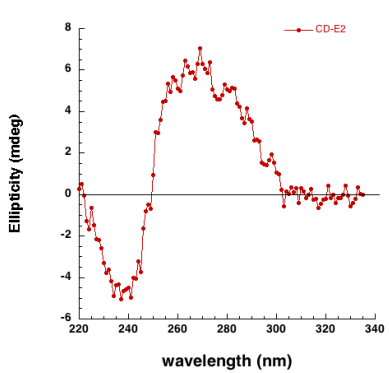
(b)



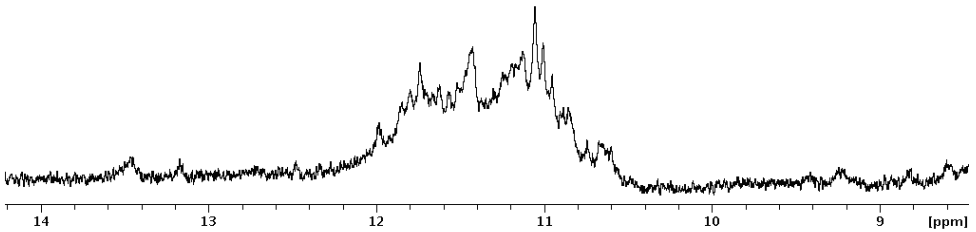
(c)



(d)



(e)



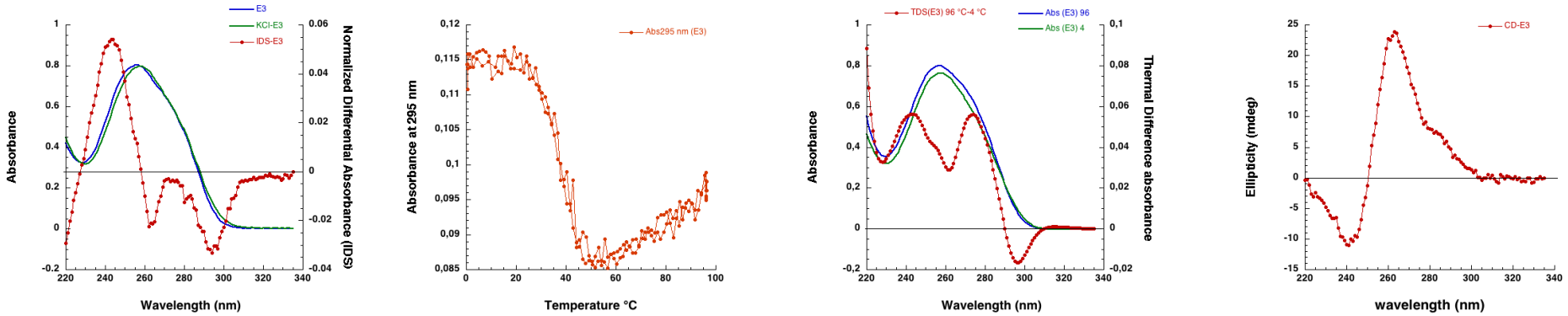
In vitro characterization of the selected candidates: **a)** Normalized Isothermal Difference Spectra (IDS) resulting from the difference between the absorbance recorded at 25°C before and after annealing in 100 mM KCl. **b)** Thermal melting transition (Tm) profiles measured at 240 and/or 295 nm (4 µM strand concentration, 10mM sodium cacodylate (pH= 7.0) and 100 mM KCl). **c)** Normalize differential spectra (TDS) in the 220 and 335 nm region (4 µM strand concentration, 10mM lithium cacodylate (pH= 7.0) and 100 mM KCl). **d)** Circular dichroism spectra (CD) (4 µM strand concentration, 10mM sodium cacodylate (pH= 7.0) and 100 mM KCl, 20°C). **e)** 1D Imino proton spectra in a 20 mM potassium phosphate buffer pH 6.9 with 70 mM KCl at 25°C. The presence of peacks between 10 and 12 ppm suggests G4 formation.

Table 3: Results interpretation of Ebola 2

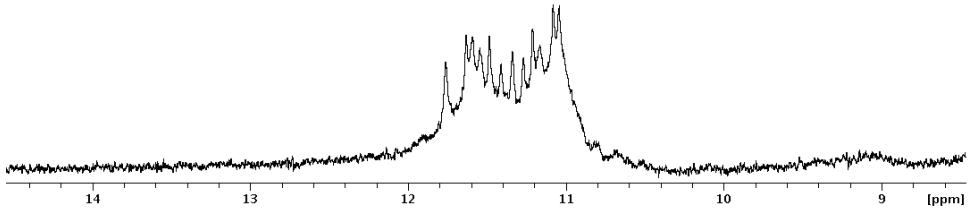
Technique	IDS	Tm	TDS	CD	NMR	**Conclusion**
Result (G4 ?)	No	No	No	No	Yes	**?? G4**

Name: Ebola 3 **Sequence:** $^{5'}GTTTTGGGGACTTGTTGTGGTGGCGGGGT^{3'}$ **Score:** 1.41

(a) (b) (c) (d)



(e)



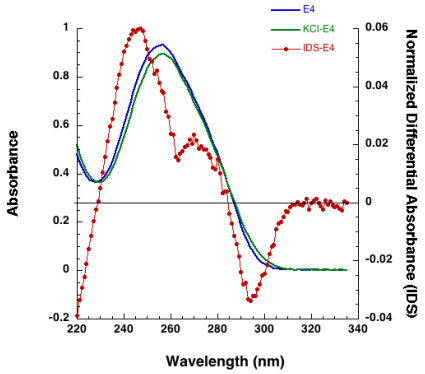
In vitro characterization of the selected candidates: **a)** Normalized Isothermal Difference Spectra (IDS) resulting from the difference between the absorbance recorded at 25°C before and after annealing in 100 mM KCl. **b)** Thermal melting transition (Tm) profiles measured at 240 and/or 295 nm (4 µM strand concentration, 10mM sodium cacodylate (pH= 7.0) and 100 mM KCl). **c)** Normalize differential spectra (TDS) in the 220 and 335 nm region (4 µM strand concentration, 10mM lithium cacodylate (pH= 7.0) and 100 mM KCl). **d)** Circular dichroism spectra (CD) (4 µM strand concentration, 10mM sodium cacodylate (pH= 7.0) and 100 mM KCl, 20°C). **e)** 1D Imino proton spectra in a 20 mM potassium phosphate buffer pH 6.9 with 70 mM KCl at 25°C. The presence of peacks between 10 and 12 ppm suggests G4 formation.

Table 4: Results interpretation of Ebola 3

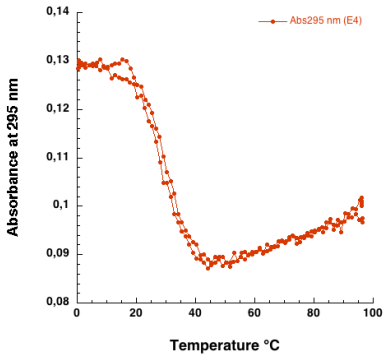
Technique	IDS	Tm	TDS	CD	NMR	**Conclusion**
Result (G4 ?)	Yes	Yes	Yes	Parallel	Yes	**G4 (Stable)**

Name: Ebola 4 **Sequence:** $^{5'}AGGGGTGGAAGGTTTATTGGGCTGGTATTG^{3'}$ **Score:** 1.23

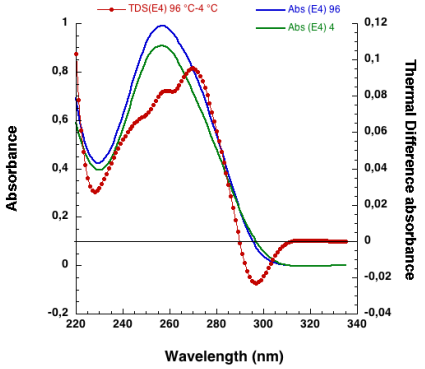
(a)



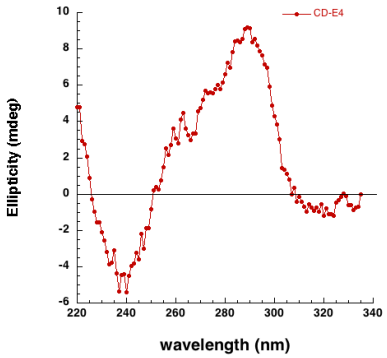
(b)



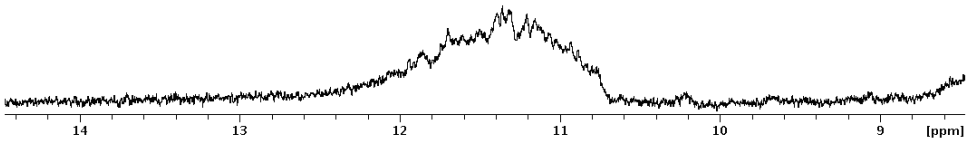
(c)



(d)



(e)



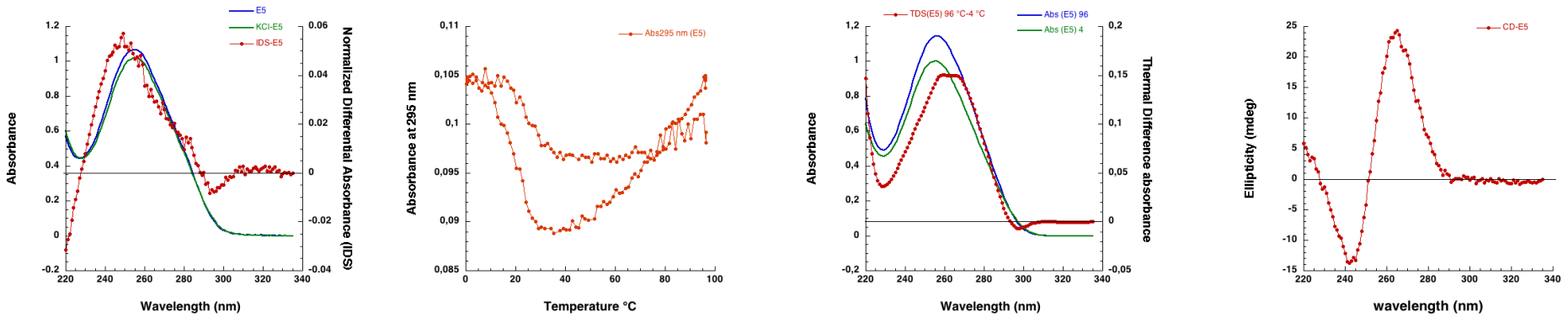
In vitro characterization of the selected candidates: **a)** Normalized Isothermal Difference Spectra (IDS) resulting from the difference between the absorbance recorded at 25°C before and after annealing in 100 mM KCl. **b)** Thermal melting transition (Tm) profiles measured at 240 and/or 295 nm (4 µM strand concentration, 10mM sodium cacodylate (pH= 7.0) and 100 mM KCl). **c)** Normalize differential spectra (TDS) in the 220 and 335 nm region (4 µM strand concentration, 10mM lithium cacodylate (pH= 7.0) and 100 mM KCl). **d)** Circular dichroism spectra (CD) (4 µM strand concentration, 10mM sodium cacodylate (pH= 7.0) and 100 mM KCl, 20°C). **e)** 1D Imino proton spectra in a 20 mM potassium phosphate buffer pH 6.9 with 70 mM KCl at 25°C. The presence of peacks between 10 and 12 ppm suggests G4 formation.

Table 5: Results interpretation of Ebola 4

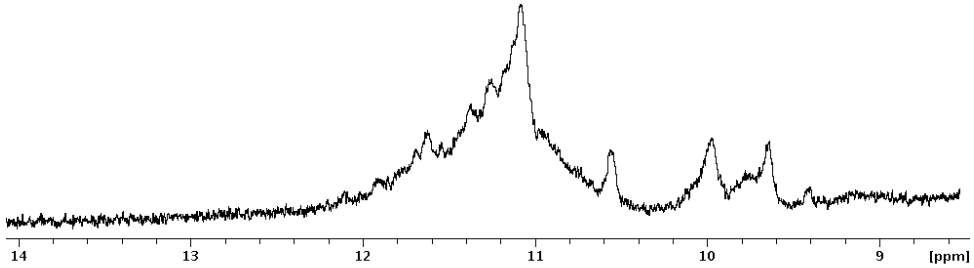
Technique	IDS	Tm	TDS	CD	NMR	**Conclusion**
Result (G4 ?)	Yes	Yes	Yes	Mixed	Yes	**G4**

Name: Ebola 5 **Sequence:** $^{5'}AGGGGTCATATGGGAGGGATTGAAGGA^{3'}$ **Score:** 1.41

(a) (b) (c) (d)



(e)



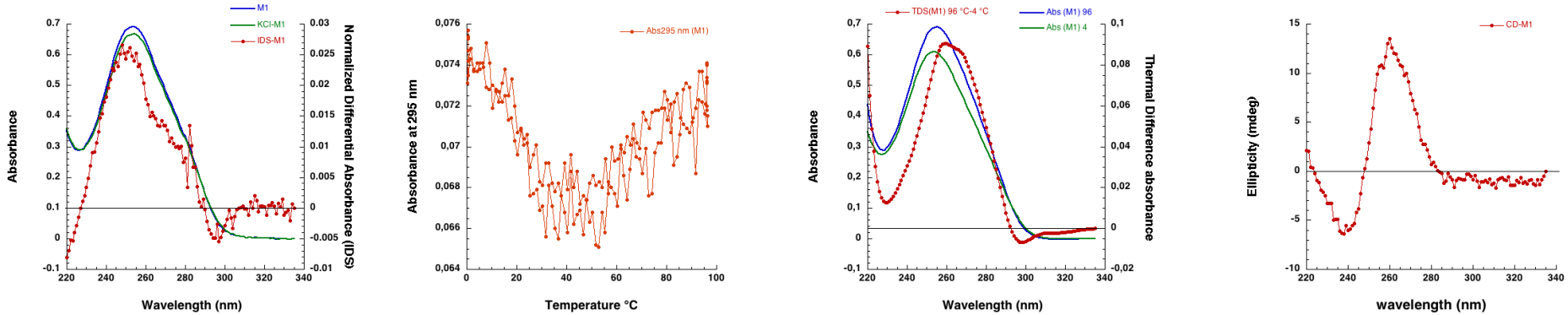
In vitro characterization of the selected candidates: **a)** Normalized Isothermal Difference Spectra (IDS) resulting from the difference between the absorbance recorded at 25°C before and after annealing in 100 mM KCl. **b)** Thermal melting transition (Tm) profiles measured at 240 and/or 295 nm (4 µM strand concentration, 10mM sodium cacodylate (pH= 7.0) and 100 mM KCl). **c)** Normalize differential spectra (TDS) in the 220 and 335 nm region (4 µM strand concentration, 10mM lithium cacodylate (pH= 7.0) and 100 mM KCl). **d)** Circular dichroism spectra (CD) (4 µM strand concentration, 10mM sodium cacodylate (pH= 7.0) and 100 mM KCl, 20°C). **e)** 1D Imino proton spectra in a 20 mM potassium phosphate buffer pH 6.9 with 70 mM KCl at 25°C. The presence of peacks between 10 and 12 ppm suggests G4 formation.

Table 6: Results interpretation of Ebola 5

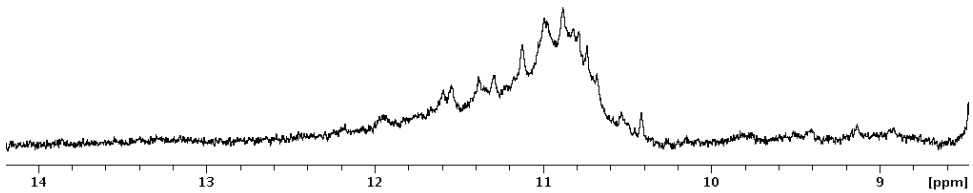
Technique	IDS	Tm	TDS	CD	NMR	**Conclusion**
Result (G4 ?)	Yes (-)	Yes	Yes (-)	Parallel	Yes	**G4**

Name: Marburg 1 **Sequence:** $^{5'}AGAGGGGGAGGATTGGGC\ ^{3'}$ **Score:** 1.83

(a) (b) (c) (d)



(e)



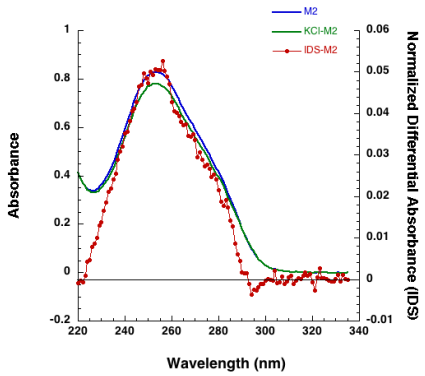
In vitro characterization of the selected candidates: **a)** Normalized Isothermal Difference Spectra (IDS) resulting from the difference between the absorbance recorded at 25°C before and after annealing in 100 mM KCl. **b)** Thermal melting transition (Tm) profiles measured at 240 and/or 295 nm (4 µM strand concentration, 10mM sodium cacodylate (pH= 7.0) and 100 mM KCl). **c)** Normalize differential spectra (TDS) in the 220 and 335 nm region (4 µM strand concentration, 10mM lithium cacodylate (pH= 7.0) and 100 mM KCl). **d)** Circular dichroism spectra (CD) (4 µM strand concentration, 10mM sodium cacodylate (pH= 7.0) and 100 mM KCl, 20°C). **e)** 1D Imino proton spectra in a 20 mM potassium phosphate buffer pH 6.9 with 70 mM KCl at 25°C. The presence of peacks between 10 and 12 ppm suggests G4 formation.

Table 7: Results interpretation of Marburg 1

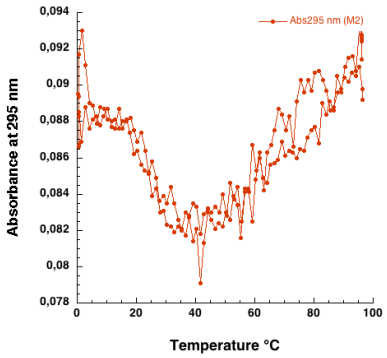
Technique	IDS	Tm	TDS	CD	NMR	**Conclusion**
Result (G4 ?)	Yes (-)	Yes (-)	Yes (-)	Parallel	Yes	**G4**

Name: Marburg 2 **Sequence:** $^{5'}CGCGGGTTGAGGAGGAGGGA^{3'}$ **Score:** 1.3

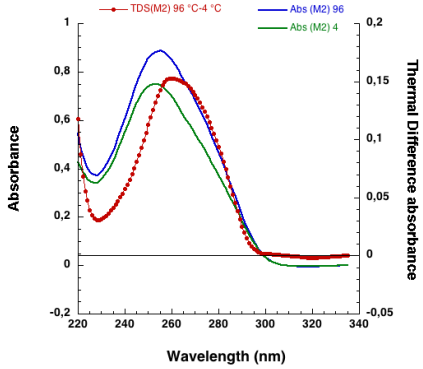
(a)



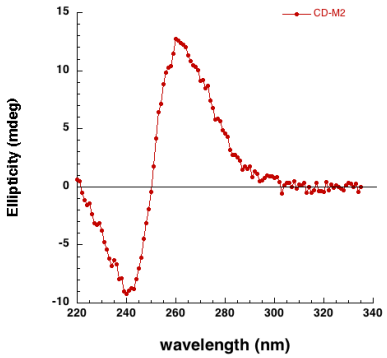
(b)



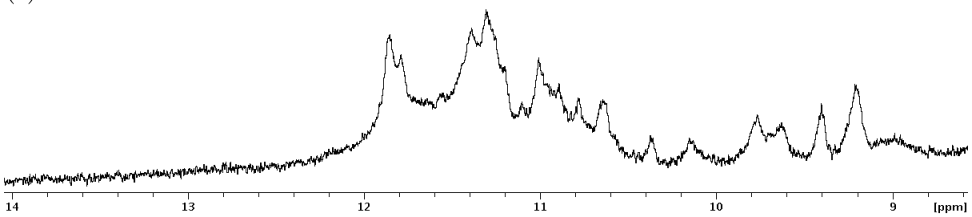
(c)



(d)



(e)



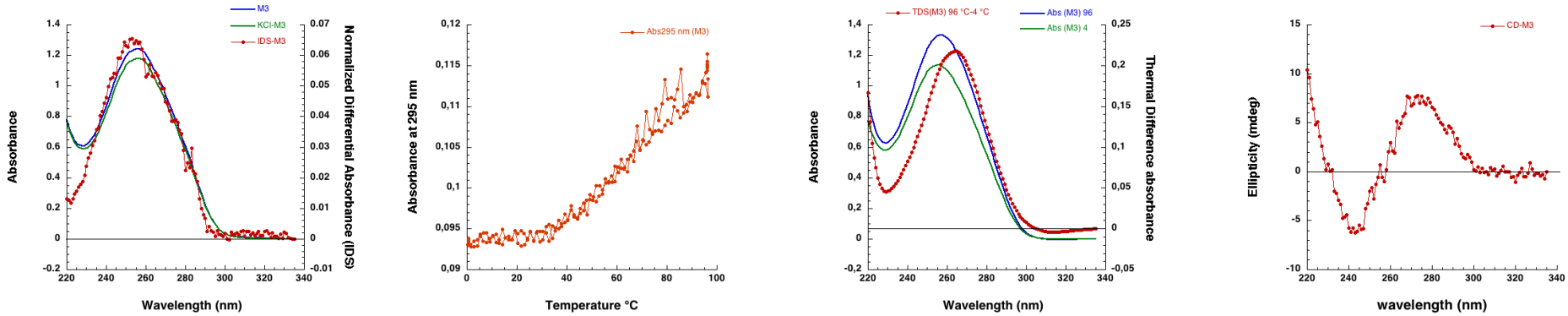
In vitro characterization of the selected candidates: **a)** Normalized Isothermal Difference Spectra (IDS) resulting from the difference between the absorbance recorded at 25°C before and after annealing in 100 mM KCl. **b)** Thermal melting transition (Tm) profiles measured at 240 and/or 295 nm (4 µM strand concentration, 10mM sodium cacodylate (pH= 7.0) and 100 mM KCl). **c)** Normalize differential spectra (TDS) in the 220 and 335 nm region (4 µM strand concentration, 10mM lithium cacodylate (pH= 7.0) and 100 mM KCl). **d)** Circular dichroism spectra (CD) (4 µM strand concentration, 10mM sodium cacodylate (pH= 7.0) and 100 mM KCl, 20°C). **e)** 1D Imino proton spectra in a 20 mM potassium phosphate buffer pH 6.9 with 70 mM KCl at 25°C. The presence of peacks between 10 and 12 ppm suggests G4 formation.

Table 8: Results interpretation of Marburg 2

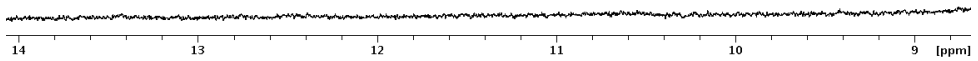
Technique	IDS	Tm	TDS	CD	NMR	**Conclusion**
Result (G4 ?)	No	Yes (-)	No	Parallel	Yes	**G4**

Name: Marburg 3 **Sequence:** $^{5'}$ *CGTGATCAGCATAAGGAGGAGGGTTCAAGT* $^{3'}$ **Score:** 0.57

(a) (b) (c) (d)



(e)



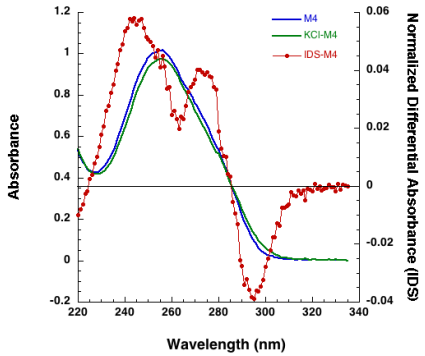
In vitro characterization of the selected candidates: **a)** Normalized Isothermal Difference Spectra (IDS) resulting from the difference between the absorbance recorded at 25°C before and after annealing in 100 mM KCl. **b)** Thermal melting transition (Tm) profiles measured at 240 and/or 295 nm (4 µM strand concentration, 10mM sodium cacodylate (pH= 7.0) and 100 mM KCl). **c)** Normalize differential spectra (TDS) in the 220 and 335 nm region (4 µM strand concentration, 10mM lithium cacodylate (pH= 7.0) and 100 mM KCl). **d)** Circular dichroism spectra (CD) (4 µM strand concentration, 10mM sodium cacodylate (pH= 7.0) and 100 mM KCl, 20°C). **e)** 1D Imino proton spectra in a 20 mM potassium phosphate buffer pH 6.9 with 70 mM KCl at 25°C. The presence of peacks between 10 and 12 ppm suggests G4 formation.

Table 9: Results interpretation of Marburg 3

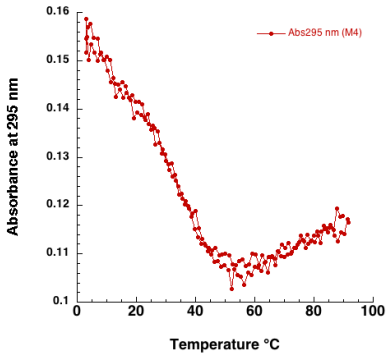
Technique	IDS	Tm	TDS	CD	NMR	**Conclusion**
Result (G4 ?)	No	No	No	No	No	**Not G4**

Name: Marburg 4 **Sequence:** $^{5'}$*CGGATGGGCTGTGGGCAGTGGTAAAGGT* $^{3'}$ **Score:** 1.04

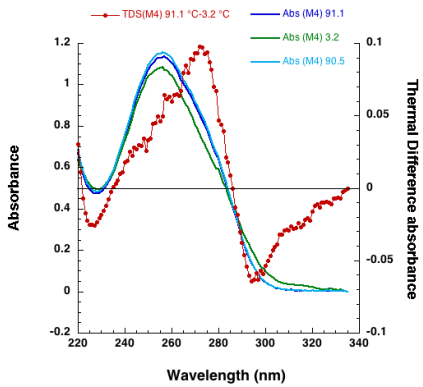
(a)



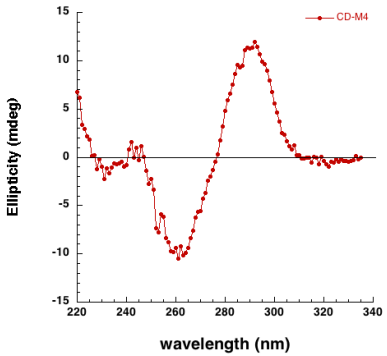
(b)



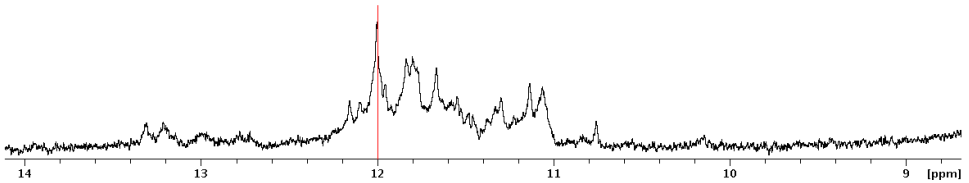
(c)



(d)



(e)



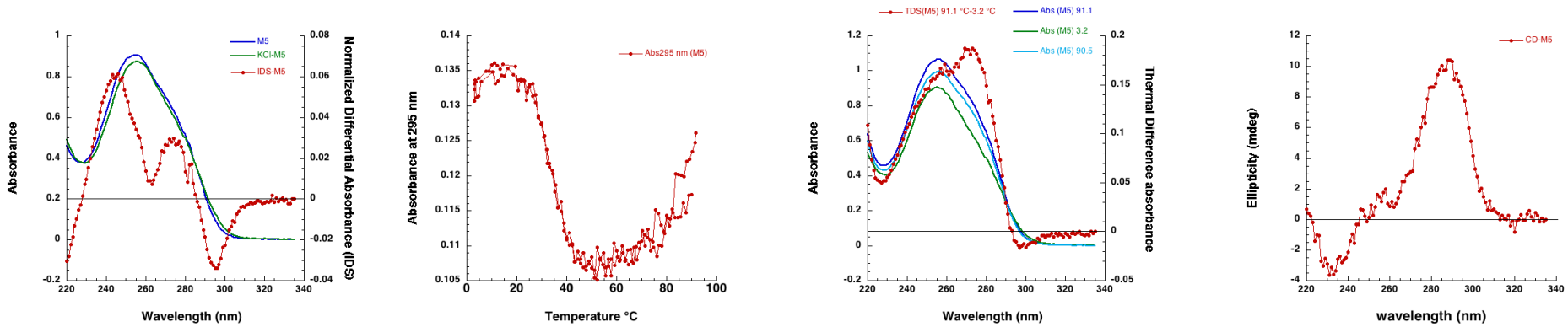
In vitro characterization of the selected candidates: **a)** Normalized Isothermal Difference Spectra (IDS) resulting from the difference between the absorbance recorded at 25°C before and after annealing in 100 mM KCl. **b)** Thermal melting transition (Tm) profiles measured at 240 and/or 295 nm (4 µM strand concentration, 10mM sodium cacodylate (pH= 7.0) and 100 mM KCl). **c)** Normalize differential spectra (TDS) in the 220 and 335 nm region (4 µM strand concentration, 10mM lithium cacodylate (pH= 7.0) and 100 mM KCl). **d)** Circular dichroism spectra (CD) (4 µM strand concentration, 10mM sodium cacodylate (pH= 7.0) and 100 mM KCl, 20°C). **e)** 1D Imino proton spectra in a 20 mM potassium phosphate buffer pH 6.9 with 70 mM KCl at 25°C. The presence of peacks between 10 and 12 ppm suggests G4 formation.

Table 10: Results interpretation of Marburg 4

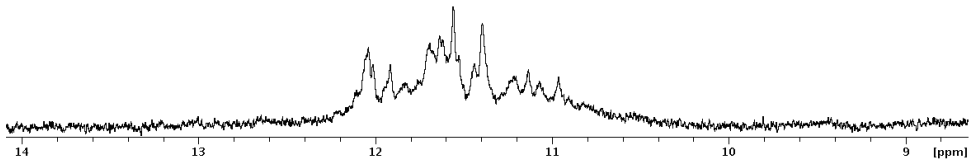
Technique	IDS	Tm	TDS	CD	NMR	**Conclusion**
Result (G4 ?)	Yes	Yes (< 37°)	Yes	Antiparallel	Yes	**G4 (UnStable)**

11 **Name:** Marburg 5 **Sequence:** $^{5'}$ *GCGTGCTTGGTTGTGGTGAGGGAGTGGGTGGC* $^{3'}$ **Score:** 1.03

(a) (b) (c) (d)



(e)



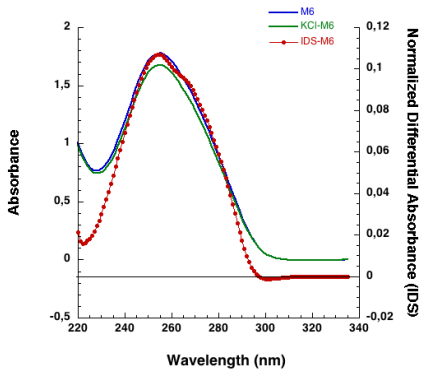
In vitro characterization of the selected candidates: **a)** Normalized Isothermal Difference Spectra (IDS) resulting from the difference between the absorbance recorded at 25°C before and after annealing in 100 mM KCl. **b)** Thermal melting transition (Tm) profiles measured at 240 and/or 295 nm (4 µM strand concentration, 10mM sodium cacodylate (pH= 7.0) and 100 mM KCl). **c)** Normalize differential spectra (TDS) in the 220 and 335 nm region (4 µM strand concentration, 10mM lithium cacodylate (pH= 7.0) and 100 mM KCl). **d)** Circular dichroism spectra (CD) (4 µM strand concentration, 10mM sodium cacodylate (pH= 7.0) and 100 mM KCl, 20°C). **e)** 1D Imino proton spectra in a 20 mM potassium phosphate buffer pH 6.9 with 70 mM KCl at 25°C. The presence of peacks between 10 and 12 ppm suggests G4 formation.

Table 11: Results interpretation of Marburg 5

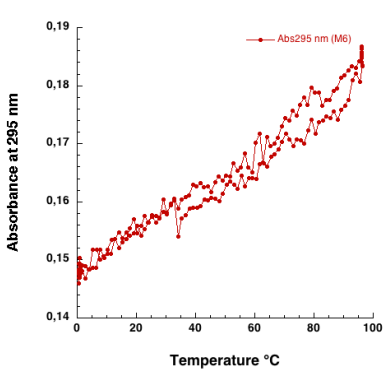
Technique	IDS	Tm	TDS	CD	NMR	**Conclusion**
Result (G4 ?)	Yes	Yes	Yes	No	Yes	**G4 (Stable)**

Name: Marburg 6 **Sequence:** $^{5'}AGAGATTGTAGTAGTGTTGTGAGGGGGCATGGACGGTTGTGCAGTTG\ ^{3'}$ **Score:** 0.79

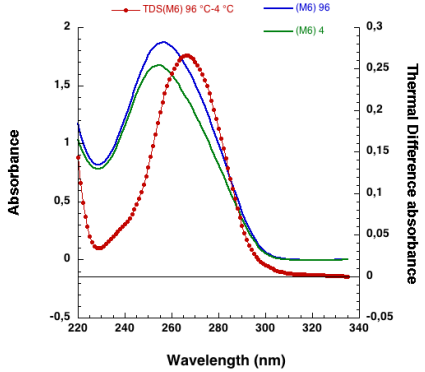
(a)



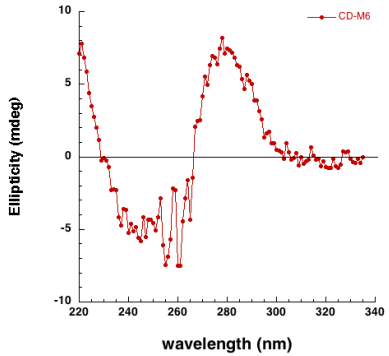
(b)



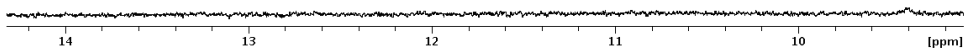
(c)



(d)



(e)



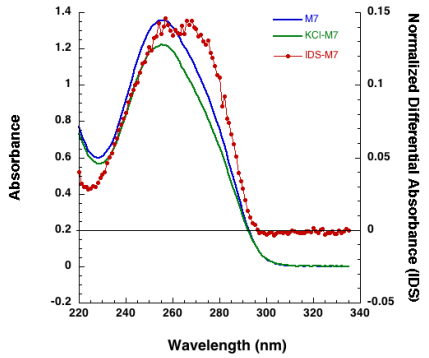
In vitro characterization of the selected candidates: **a)** Normalized Isothermal Difference Spectra (IDS) resulting from the difference between the absorbance recorded at 25°C before and after annealing in 100 mM KCl. **b)** Thermal melting transition (Tm) profiles measured at 240 and/or 295 nm (4 µM strand concentration, 10mM sodium cacodylate (pH= 7.0) and 100 mM KCl). **c)** Normalize differential spectra (TDS) in the 220 and 335 nm region (4 µM strand concentration, 10mM lithium cacodylate (pH= 7.0) and 100 mM KCl). **d)** Circular dichroism spectra (CD) (4 µM strand concentration, 10mM sodium cacodylate (pH= 7.0) and 100 mM KCl, 20°C). **e)** 1D Imino proton spectra in a 20 mM potassium phosphate buffer pH 6.9 with 70 mM KCl at 25°C. The presence of peacks between 10 and 12 ppm suggests G4 formation.

Table 12: Results interpretation of Marburg 6

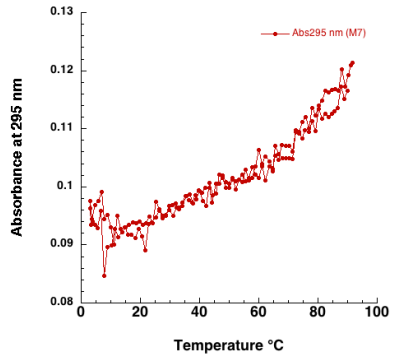
Technique	IDS	Tm	TDS	CD	NMR	**Conclusion**
Result (G4 ?)	No	No	No	No	No	**Not G4**

13 **Name:** Marburg 7 **Sequence:** $^{5'}$*TGTTGGAAGCAGGGTTGTTTTCGAGGGGGCACTGGT* $^{3'}$ **Score:** 1.03

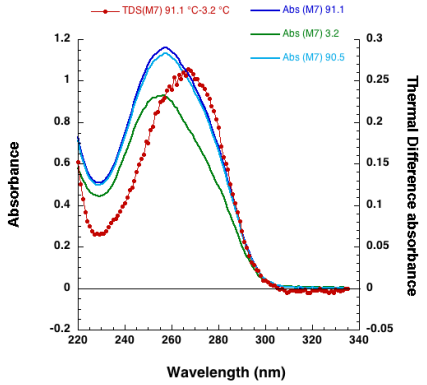
(a)



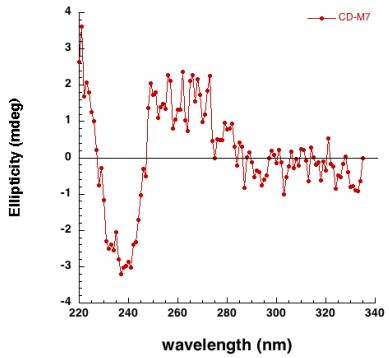
(b)



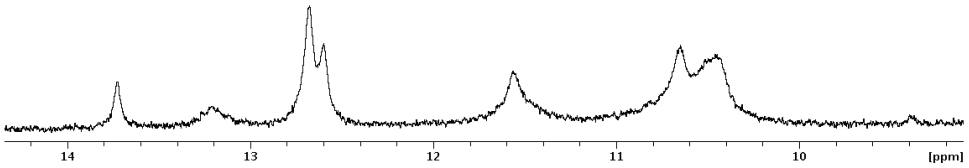
(c)



(d)



(e)



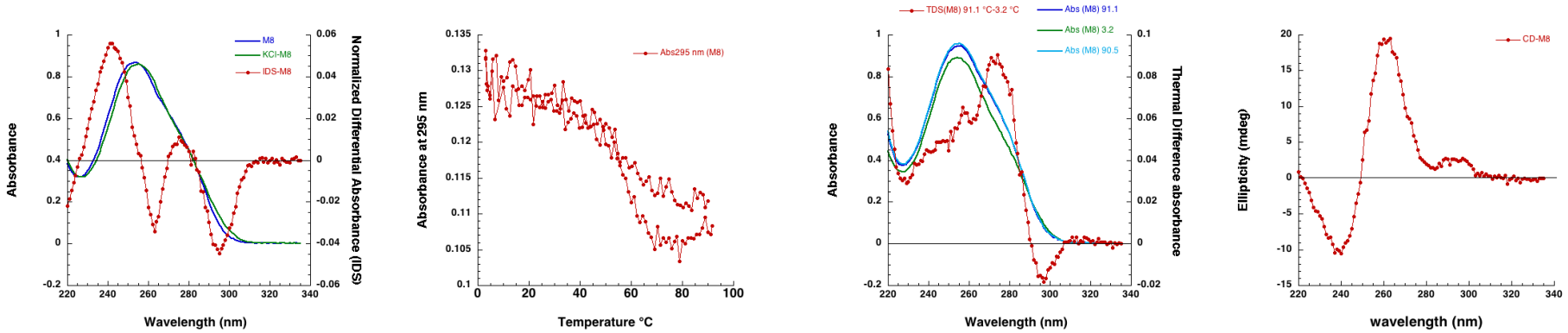
In vitro characterization of the selected candidates: **a)** Normalized Isothermal Difference Spectra (IDS) resulting from the difference between the absorbance recorded at 25°C before and after annealing in 100 mM KCl. **b)** Thermal melting transition (Tm) profiles measured at 240 and/or 295 nm (4 µM strand concentration, 10mM sodium cacodylate (pH= 7.0) and 100 mM KCl). **c)** Normalize differential spectra (TDS) in the 220 and 335 nm region (4 µM strand concentration, 10mM lithium cacodylate (pH= 7.0) and 100 mM KCl). **d)** Circular dichroism spectra (CD) (4 µM strand concentration, 10mM sodium cacodylate (pH= 7.0) and 100 mM KCl, 20°C). **e)** 1D Imino proton spectra in a 20 mM potassium phosphate buffer pH 6.9 with 70 mM KCl at 25°C. The presence of peacks between 10 and 12 ppm suggests G4 formation.

Table 13: Results interpretation of Marburg 7

Technique	IDS	Tm	TDS	CD	NMR	**Conclusion**
Result (G4 ?)	No	No	No	No	No	**Not G4**

Name: Marburg 8 **Sequence:** $^{5'}TGGGGGTGGGGGAGGGACTGGTGGA\ ^{3'}$ **Score:** 2.24

(a) (b) (c) (d)



(e)

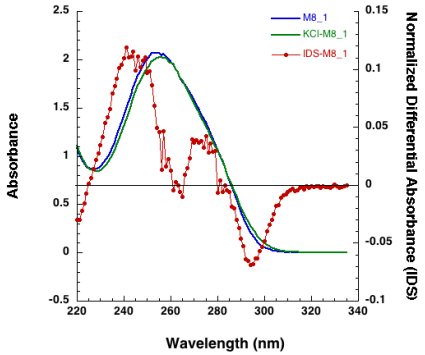
In vitro characterization of the selected candidates: **a)** Normalized Isothermal Difference Spectra (IDS) resulting from the difference between the absorbance recorded at 25°C before and after annealing in 100 mM KCl. **b)** Thermal melting transition (Tm) profiles measured at 240 and/or 295 nm (4 µM strand concentration, 10mM sodium cacodylate (pH= 7.0) and 100 mM KCl). **c)** Normalize differential spectra (TDS) in the 220 and 335 nm region (4 µM strand concentration, 10mM lithium cacodylate (pH= 7.0) and 100 mM KCl). **d)** Circular dichroism spectra (CD) (4 µM strand concentration, 10mM sodium cacodylate (pH= 7.0) and 100 mM KCl, 20°C). **e)** 1D Imino proton spectra in a 20 mM potassium phosphate buffer pH 6.9 with 70 mM KCl at 25°C. The presence of peacks between 10 and 12 ppm suggests G4 formation.

Table 14: Results interpretation of Marburg 8

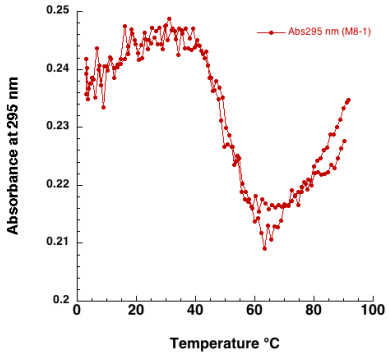
Technique	IDS	Tm	TDS	CD	NMR	**Conclusion**
Result (G4 ?)	Yes	Yes	Yes	Parallel	Not done	**G4 (Stable)**

Name: Marburg 8-1 **Sequence:** $^{5'}$*CAAGATGTTGTGCAGTCGAGTTGGGGGTGGGGGAGGGACTGGTGGAATAC* $^{3'}$ **Score:** 1.18

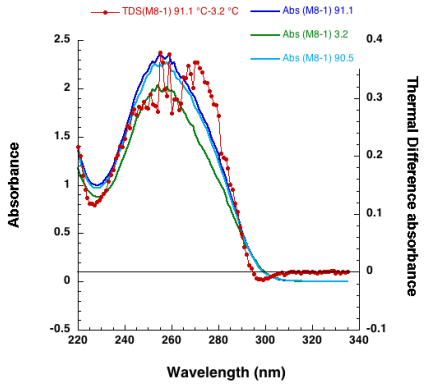
(a)



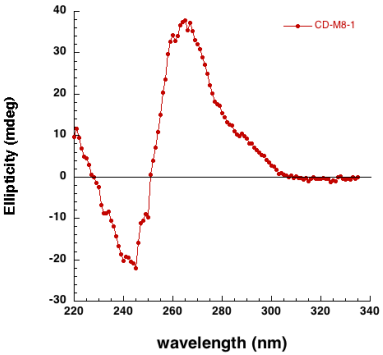
(b)



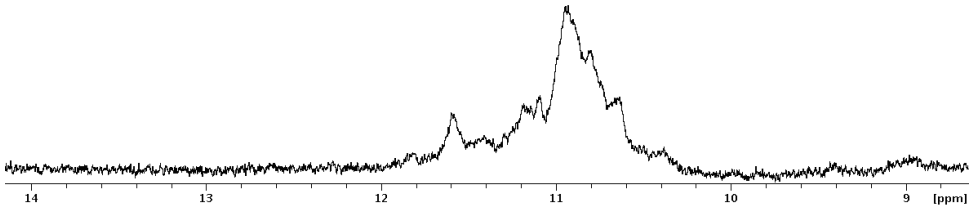
(c)



(d)



(e)



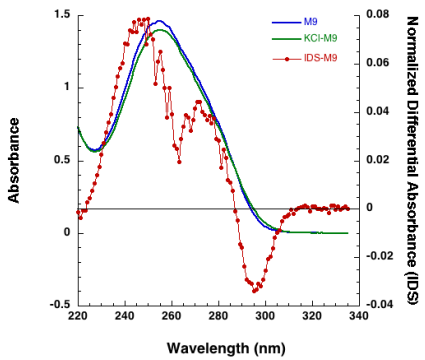
In vitro characterization of the selected candidates: **a)** Normalized Isothermal Difference Spectra (IDS) resulting from the difference between the absorbance recorded at 25°C before and after annealing in 100 mM KCl. **b)** Thermal melting transition (Tm) profiles measured at 240 and/or 295 nm (4 µM strand concentration, 10mM sodium cacodylate (pH= 7.0) and 100 mM KCl). **c)** Normalize differential spectra (TDS) in the 220 and 335 nm region (4 µM strand concentration, 10mM lithium cacodylate (pH= 7.0) and 100 mM KCl). **d)** Circular dichroism spectra (CD) (4 µM strand concentration, 10mM sodium cacodylate (pH= 7.0) and 100 mM KCl, 20°C). **e)** 1D Imino proton spectra in a 20 mM potassium phosphate buffer pH 6.9 with 70 mM KCl at 25°C. The presence of peacks between 10 and 12 ppm suggests G4 formation.

Table 15: Results interpretation of Marburg 8-1

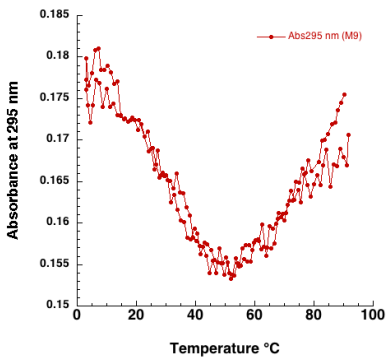
Technique	IDS	Tm	TDS	CD	NMR	**Conclusion**
Result (G4 ?)	Yes	Yes	No	Parallel	Yes	**G4 (Stable)**

Name: Marburg 9 **Sequence:** $^{5'}AGAGGGGACTGGTTGGGGTCTGGGTGGTAAATGGTGGA^{3'}$ **Score:** 1.47

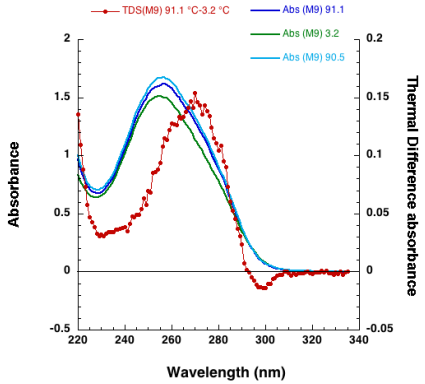
(a)



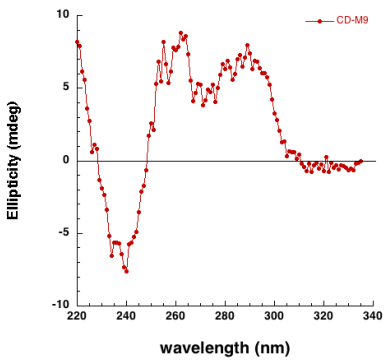
(b)



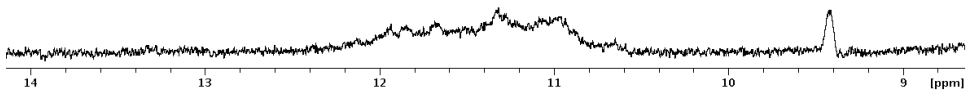
(c)



(d)



(e)



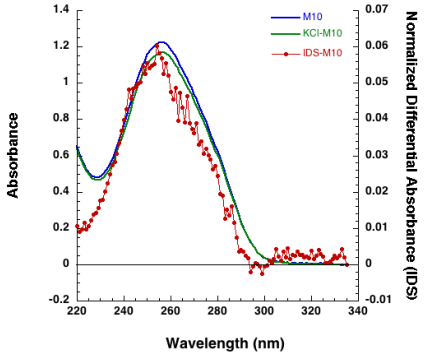
In vitro characterization of the selected candidates: **a)** Normalized Isothermal Difference Spectra (IDS) resulting from the difference between the absorbance recorded at 25°C before and after annealing in 100 mM KCl. **b)** Thermal melting transition (Tm) profiles measured at 240 and/or 295 nm (4 µM strand concentration, 10mM sodium cacodylate (pH= 7.0) and 100 mM KCl). **c)** Normalize differential spectra (TDS) in the 220 and 335 nm region (4 µM strand concentration, 10mM lithium cacodylate (pH= 7.0) and 100 mM KCl). **d)** Circular dichroism spectra (CD) (4 µM strand concentration, 10mM sodium cacodylate (pH= 7.0) and 100 mM KCl, 20°C). **e)** 1D Imino proton spectra in a 20 mM potassium phosphate buffer pH 6.9 with 70 mM KCl at 25°C. The presence of peacks between 10 and 12 ppm suggests G4 formation.

Table 16: Results interpretation of Marburg 9

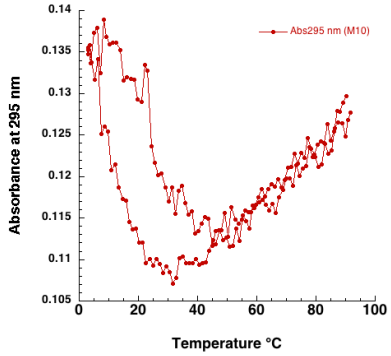
Technique	IDS	Tm	TDS	CD	NMR	**Conclusion**
Result (G4 ?)	Yes	Yes	Yes	Mixed	Yes	**G4 (Stable)**

Name: Marburg 10 **Sequence:** $^{5'}ATGATTTGTGTAATGGGGTTTGGAGGG\ ^{3'}$ **Score:** 1.19

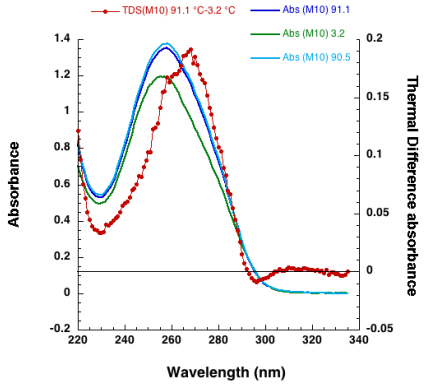
(a)



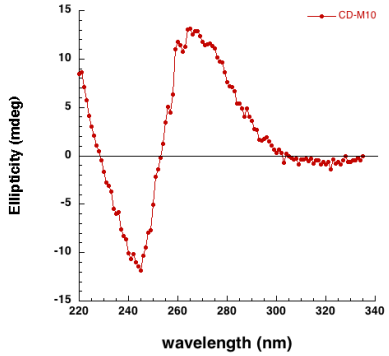
(b)



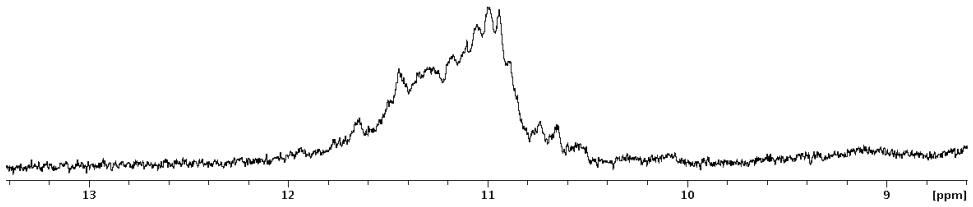
(c)



(d)



(e)



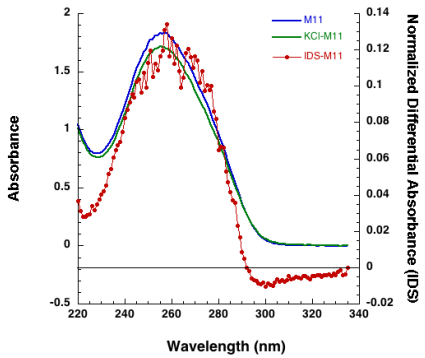
In vitro characterization of the selected candidates: **a)** Normalized Isothermal Difference Spectra (IDS) resulting from the difference between the absorbance recorded at 25°C before and after annealing in 100 mM KCl. **b)** Thermal melting transition (Tm) profiles measured at 240 and/or 295 nm (4 µM strand concentration, 10mM sodium cacodylate (pH= 7.0) and 100 mM KCl). **c)** Normalize differential spectra (TDS) in the 220 and 335 nm region (4 µM strand concentration, 10mM lithium cacodylate (pH= 7.0) and 100 mM KCl). **d)** Circular dichroism spectra (CD) (4 µM strand concentration, 10mM sodium cacodylate (pH= 7.0) and 100 mM KCl, 20°C). **e)** 1D Imino proton spectra in a 20 mM potassium phosphate buffer pH 6.9 with 70 mM KCl at 25°C. The presence of peacks between 10 and 12 ppm suggests G4 formation.

Table 17: Results interpretation of Marburg 10

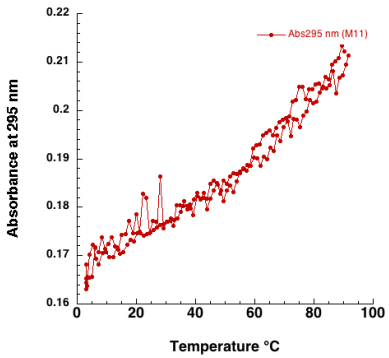
Technique	IDS	Tm	TDS	CD	NMR	**Conclusion**
Result (G4 ?)	No	Yes	No	No	Yes	**G4**

Name: Marburg 11 **Sequence:** $^{5'}$ *GCACATGTCCTTTGGGGGAGGAGGGGGTTTCGATAAGTTGA* $^{3'}$ **Score:** 1.02

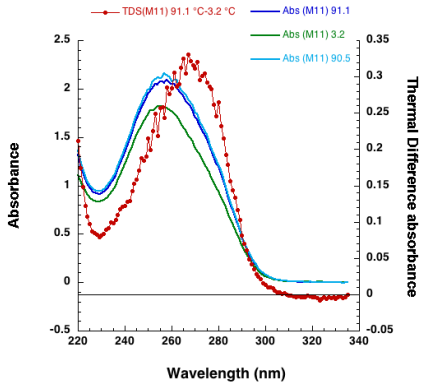
(a)



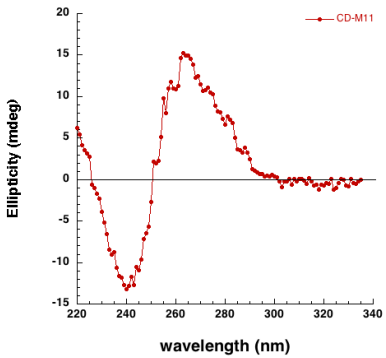
(b)



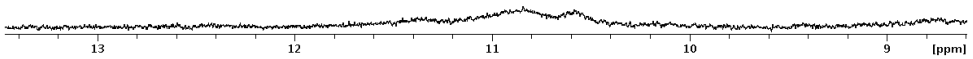
(c)



(d)



(e)



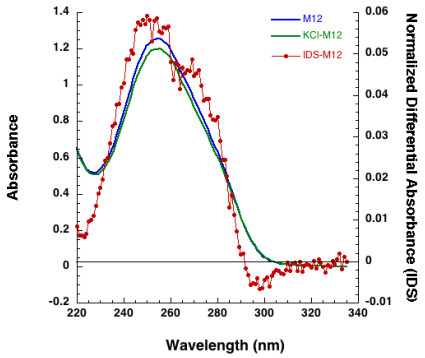
In vitro characterization of the selected candidates: **a)** Normalized Isothermal Difference Spectra (IDS) resulting from the difference between the absorbance recorded at 25°C before and after annealing in 100 mM KCl. **b)** Thermal melting transition (Tm) profiles measured at 240 and/or 295 nm (4 µM strand concentration, 10mM sodium cacodylate (pH= 7.0) and 100 mM KCl). **c)** Normalize differential spectra (TDS) in the 220 and 335 nm region (4 µM strand concentration, 10mM lithium cacodylate (pH= 7.0) and 100 mM KCl). **d)** Circular dichroism spectra (CD) (4 µM strand concentration, 10mM sodium cacodylate (pH= 7.0) and 100 mM KCl, 20°C). **e)** 1D Imino proton spectra in a 20 mM potassium phosphate buffer pH 6.9 with 70 mM KCl at 25°C. The presence of peacks between 10 and 12 ppm suggests G4 formation.

Table 18: Results interpretation of Marburg 11

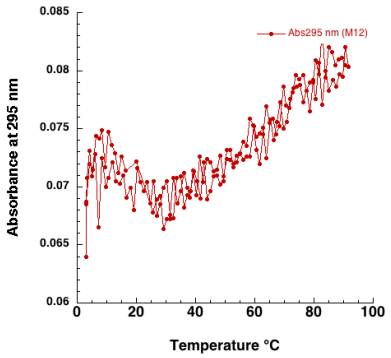
Technique	IDS	Tm	TDS	CD	NMR	**Conclusion**
Result (G4 ?)	No	No	No	No	No	**Not G4**

Name: Marburg 12 **Sequence:** $^{5'}AGTGTTGGGGGCTGGACAGTGAAGTGGGGA^{3'}$ **Score:** 1.43

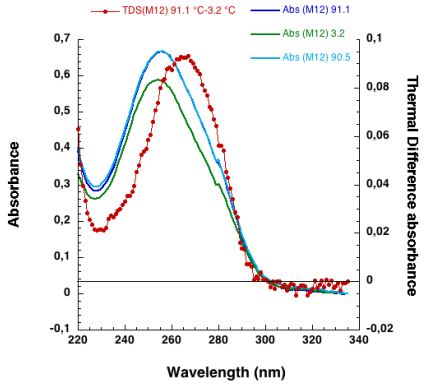
(a)



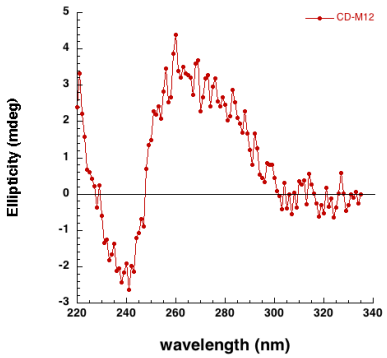
(b)



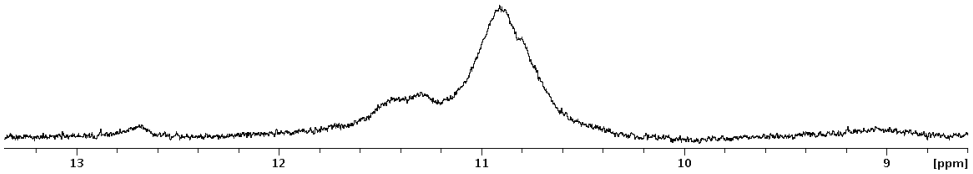
(c)



(d)



(e)



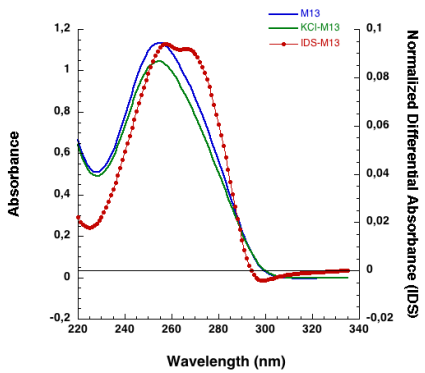
In vitro characterization of the selected candidates: **a)** Normalized Isothermal Difference Spectra (IDS) resulting from the difference between the absorbance recorded at 25°C before and after annealing in 100 mM KCl. **b)** Thermal melting transition (Tm) profiles measured at 240 and/or 295 nm (4 µM strand concentration, 10mM sodium cacodylate (pH= 7.0) and 100 mM KCl). **c)** Normalize differential spectra (TDS) in the 220 and 335 nm region (4 µM strand concentration, 10mM lithium cacodylate (pH= 7.0) and 100 mM KCl). **d)** Circular dichroism spectra (CD) (4 µM strand concentration, 10mM sodium cacodylate (pH= 7.0) and 100 mM KCl, 20°C). **e)** 1D Imino proton spectra in a 20 mM potassium phosphate buffer pH 6.9 with 70 mM KCl at 25°C. The presence of peacks between 10 and 12 ppm suggests G4 formation.

Table 19: Results interpretation of Marburg 12

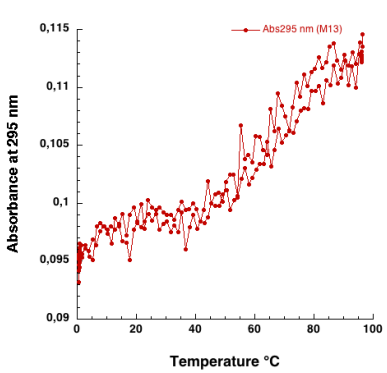
Technique	IDS	Tm	TDS	CD	NMR	**Conclusion**
Result (G4 ?)	Yes(-)	No	No	No	Yes	**??G4**

Name: Marburg 13 **Sequence:** $^{5'}$ *TGGGGTTTCTAGTGGAAGTCAGGAGGA* $^{3'}$ **Score:** 1.04

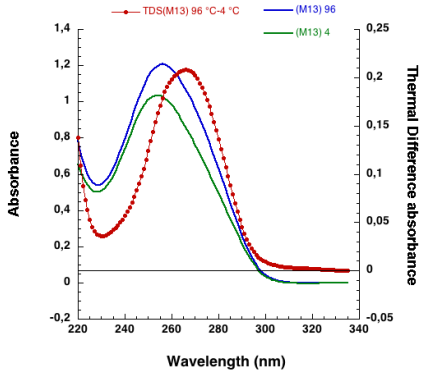
(a)



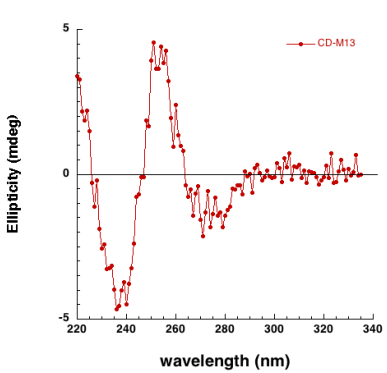
(b)



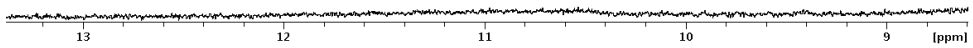
(c)



(d)



(e)



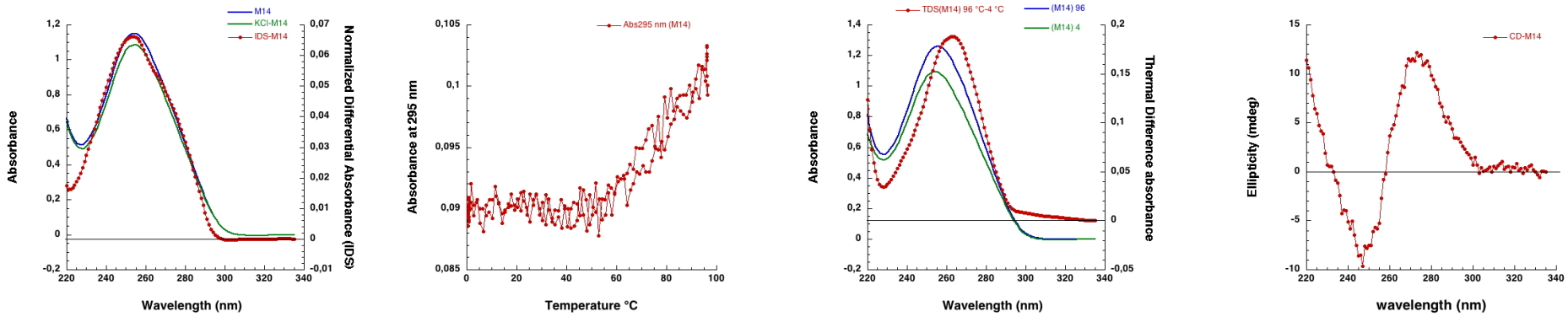
In vitro characterization of the selected candidates: **a)** Normalized Isothermal Difference Spectra (IDS) resulting from the difference between the absorbance recorded at 25°C before and after annealing in 100 mM KCl. **b)** Thermal melting transition (Tm) profiles measured at 240 and/or 295 nm (4 µM strand concentration, 10mM sodium cacodylate (pH= 7.0) and 100 mM KCl). **c)** Normalize differential spectra (TDS) in the 220 and 335 nm region (4 µM strand concentration, 10mM lithium cacodylate (pH= 7.0) and 100 mM KCl). **d)** Circular dichroism spectra (CD) (4 µM strand concentration, 10mM sodium cacodylate (pH= 7.0) and 100 mM KCl, 20°C). **e)** 1D Imino proton spectra in a 20 mM potassium phosphate buffer pH 6.9 with 70 mM KCl at 25°C. The presence of peacks between 10 and 12 ppm suggests G4 formation.

Table 20: Results interpretation of Marburg 13

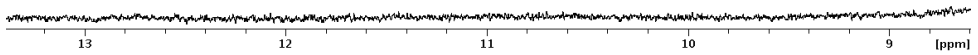
Technique	IDS	Tm	TDS	CD	NMR	**Conclusion**
Result (G4 ?)	No	No	No	No	No	**Not G4**

Name: Marburg 14 **Sequence:** $^{5'}$ *AGGGGAAAACGCTATAGTGAGAGGTGCGAG* $^{3'}$ **score:** 0.8

(a) (b) (c) (d)



(e)



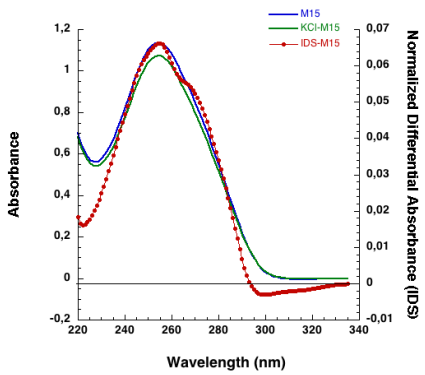
In vitro characterization of the selected candidates: **a)** Normalized Isothermal Difference Spectra (IDS) resulting from the difference between the absorbance recorded at 25°C before and after annealing in 100 mM KCl. **b)** Thermal melting transition (Tm) profiles measured at 240 and/or 295 nm (4 µM strand concentration, 10mM sodium cacodylate (pH= 7.0) and 100 mM KCl). **c)** Normalize differential spectra (TDS) in the 220 and 335 nm region (4 µM strand concentration, 10mM lithium cacodylate (pH= 7.0) and 100 mM KCl). **d)** Circular dichroism spectra (CD) (4 µM strand concentration, 10mM sodium cacodylate (pH= 7.0) and 100 mM KCl, 20°C). **e)** 1D Imino proton spectra in a 20 mM potassium phosphate buffer pH 6.9 with 70 mM KCl at 25°C. The presence of peacks between 10 and 12 ppm suggests G4 formation.

Table 21: Results interpretation of Marburg 14

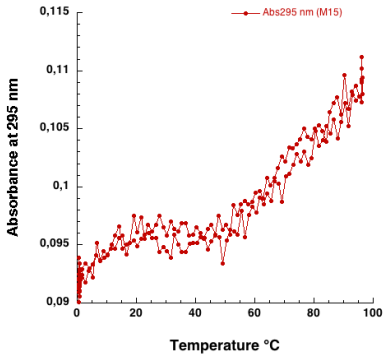
Technique	IDS	TM	TDS	CD	RMN	G4
Result (G4 ?)	No	No	No	No	No	**Not G4**

Name: Marburg 15 **Sequence:** $^{5'}$ *TGGGTAAATACGCAGGGGGAGGTCAAGCTG* $^{3'}$ **Score:** 1.07

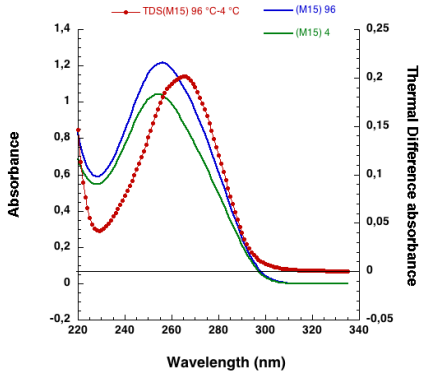
(a)



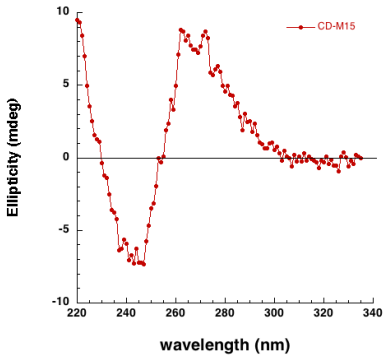
(b)



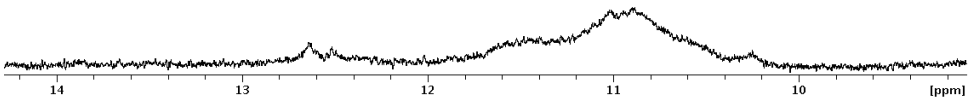
(c)



(d)



(e)



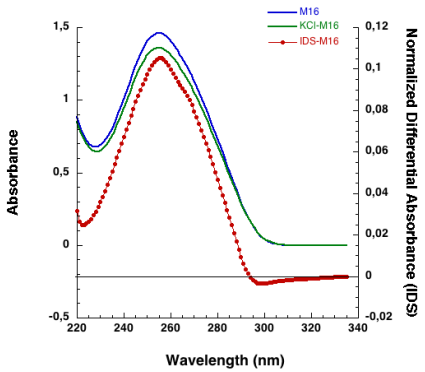
In vitro characterization of the selected candidates: **a)** Normalized Isothermal Difference Spectra (IDS) resulting from the difference between the absorbance recorded at 25°C before and after annealing in 100 mM KCl. **b)** Thermal melting transition (Tm) profiles measured at 240 and/or 295 nm (4 µM strand concentration, 10mM sodium cacodylate (pH= 7.0) and 100 mM KCl). **c)** Normalize differential spectra (TDS) in the 220 and 335 nm region (4 µM strand concentration, 10mM lithium cacodylate (pH= 7.0) and 100 mM KCl). **d)** Circular dichroism spectra (CD) (4 µM strand concentration, 10mM sodium cacodylate (pH= 7.0) and 100 mM KCl, 20°C). **e)** 1D Imino proton spectra in a 20 mM potassium phosphate buffer pH 6.9 with 70 mM KCl at 25°C. The presence of peacks between 10 and 12 ppm suggests G4 formation.

Table 22: Results interpretation of Marburg 15

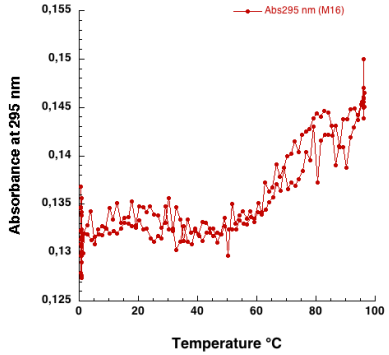
Technique	IDS	Tm	TDS	CD	NMR	**Conclusion**
Result (G4 ?)	No	No	No	No	No (faible)	**Not G4**

Name: Marburg 16 **Sequence:** $^{5'}$ *TGGTTTGAGATGGGGAGGAGCCATGTATGTACGCAG* $^{3'}$ **Score:** 0.69

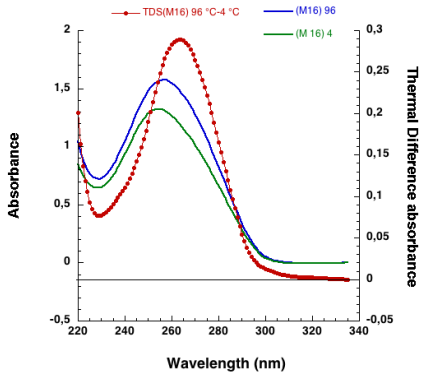
(a)



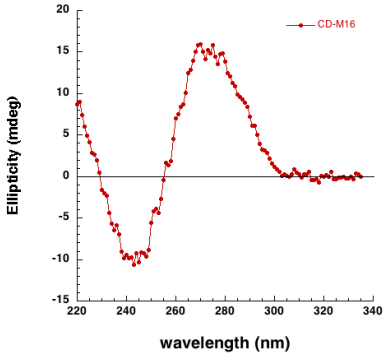
(b)



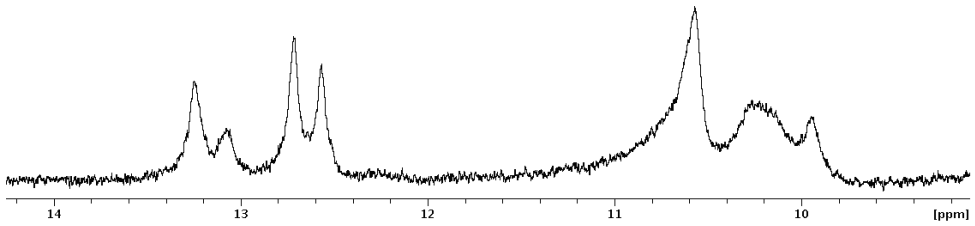
(c)



(d)



(e)



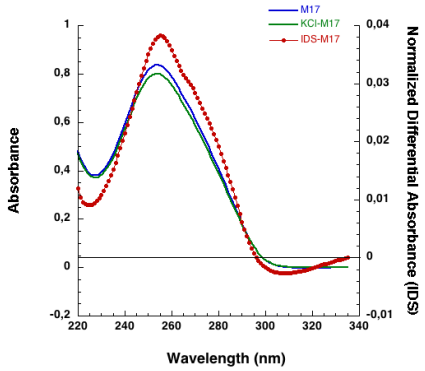
In vitro characterization of the selected candidates: **a)** Normalized Isothermal Difference Spectra (IDS) resulting from the difference between the absorbance recorded at 25°C before and after annealing in 100 mM KCl. **b)** Thermal melting transition (Tm) profiles measured at 240 and/or 295 nm (4 µM strand concentration, 10mM sodium cacodylate (pH= 7.0) and 100 mM KCl). **c)** Normalize differential spectra (TDS) in the 220 and 335 nm region (4 µM strand concentration, 10mM lithium cacodylate (pH= 7.0) and 100 mM KCl). **d)** Circular dichroism spectra (CD) (4 µM strand concentration, 10mM sodium cacodylate (pH= 7.0) and 100 mM KCl, 20°C). **e)** 1D Imino proton spectra in a 20 mM potassium phosphate buffer pH 6.9 with 70 mM KCl at 25°C. The presence of peacks between 10 and 12 ppm suggests G4 formation.

Table 23: Results interpretation of Marburg 16

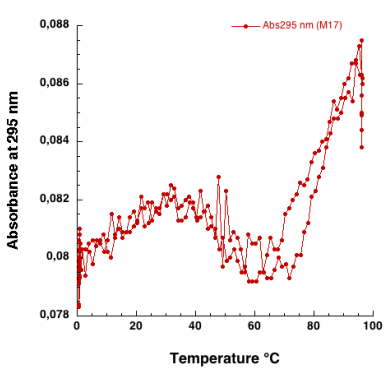
Technique	IDS	Tm	TDS	CD	NMR	**Conclusion**
Result (G4 ?)	No	No	No	No	??	**Not G4**

Name: Marburg 17 **Sequence:** *GCTTGGCCGAAGGAGAAGGAAGTGGTG* $^{3'}$ **score:** 0.59

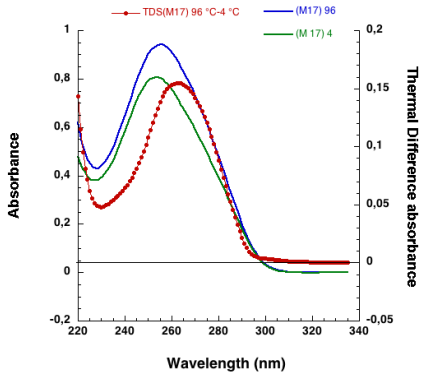
(a)



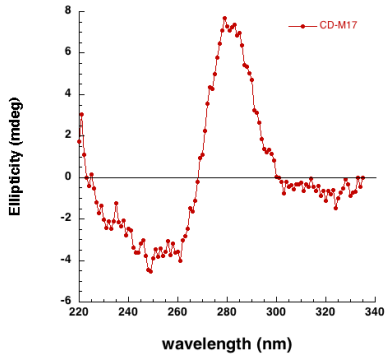
(b)



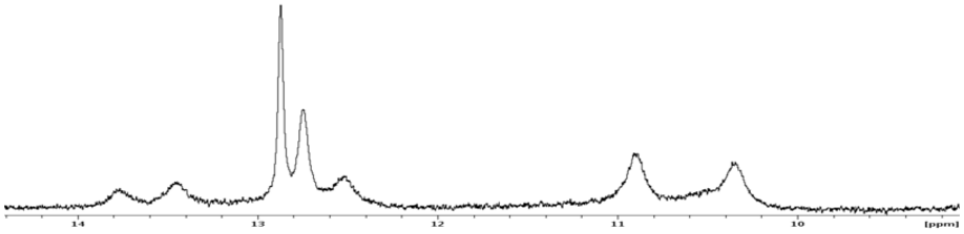
(c)



(d)



(e)



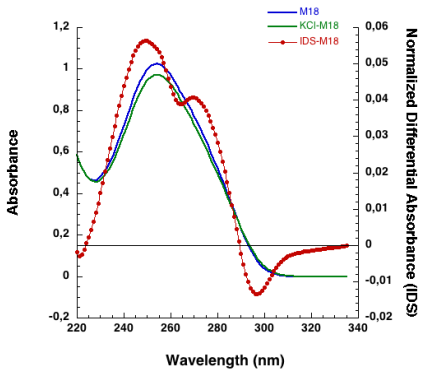
In vitro characterization of the selected candidates: **a)** Normalized Isothermal Difference Spectra (IDS) resulting from the difference between the absorbance recorded at 25°C before and after annealing in 100 mM KCl. **b)** Thermal melting transition (Tm) profiles measured at 240 and/or 295 nm (4 µM strand concentration, 10mM sodium cacodylate (pH= 7.0) and 100 mM KCl). **c)** Normalize differential spectra (TDS) in the 220 and 335 nm region (4 µM strand concentration, 10mM lithium cacodylate (pH= 7.0) and 100 mM KCl). **d)** Circular dichroism spectra (CD) (4 µM strand concentration, 10mM sodium cacodylate (pH= 7.0) and 100 mM KCl, 20°C). **e)** 1D Imino proton spectra in a 20 mM potassium phosphate buffer pH 6.9 with 70 mM KCl at 25°C. The presence of peacks between 10 and 12 ppm suggests G4 formation.

Table 24: Results interpretation of Marburg 17

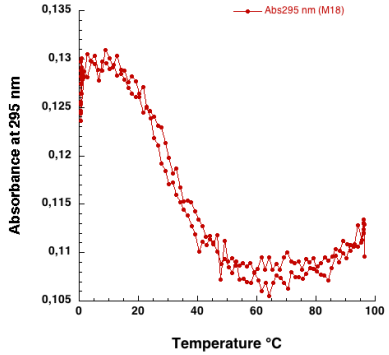
Technique	IDS	TM	TDS	CD	RMN	G4
Result (G4 ?)	No	No	No	No	??	**Not G4**

Name: Marburg 18 **Sequence:** $^{5'}$ *TTGGCCGAAGGGGAAGGAAGTGGTGCTCG* $^{3'}$ **Score:** 0.9

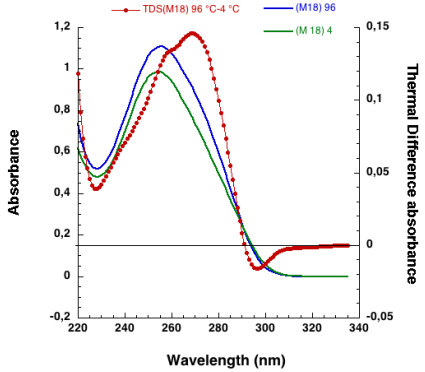
(a)



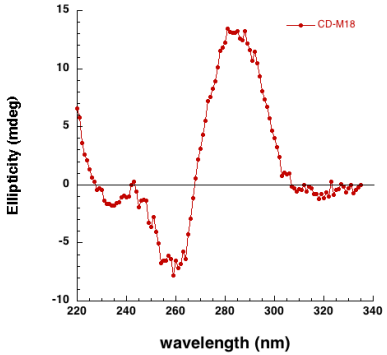
(b)



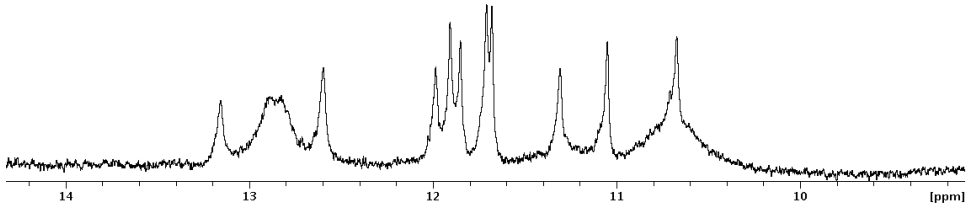
(c)



(d)



(e)



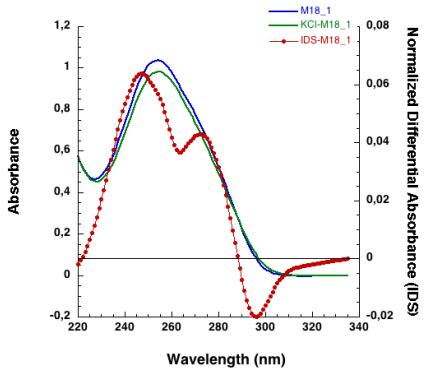
In vitro characterization of the selected candidates: **a)** Normalized Isothermal Difference Spectra (IDS) resulting from the difference between the absorbance recorded at 25°C before and after annealing in 100 mM KCl. **b)** Thermal melting transition (Tm) profiles measured at 240 and/or 295 nm (4 µM strand concentration, 10mM sodium cacodylate (pH= 7.0) and 100 mM KCl). **c)** Normalize differential spectra (TDS) in the 220 and 335 nm region (4 µM strand concentration, 10mM lithium cacodylate (pH= 7.0) and 100 mM KCl). **d)** Circular dichroism spectra (CD) (4 µM strand concentration, 10mM sodium cacodylate (pH= 7.0) and 100 mM KCl, 20°C). **e)** 1D Imino proton spectra in a 20 mM potassium phosphate buffer pH 6.9 with 70 mM KCl at 25°C. The presence of peacks between 10 and 12 ppm suggests G4 formation.

Table 25: Results interpretation of Marburg 18

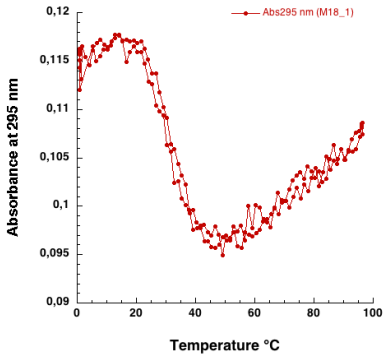
Technique	IDS	Tm	TDS	CD	NMR	**Conclusion**
Result (G4 ?)	Yes	Yes	Yes	Antiparallel	Yes	**G4 (Stable)**

Name: Marburg 18-1 **Sequence:** $^{5'}$ *TGGCTGAAGGGGAAGGAAGTGGTGCTCGGT* $^{3'}$ **Score:** 1.07

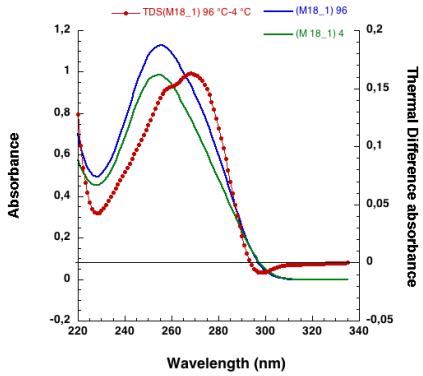
(a)



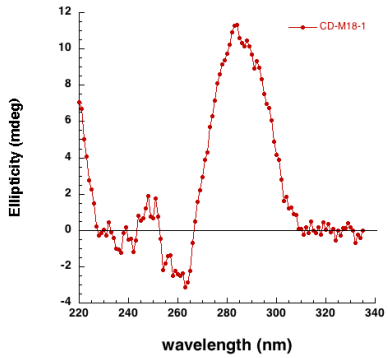
(b)



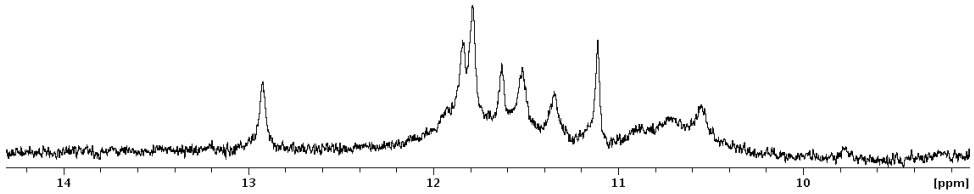
(c)



(d)



(e)



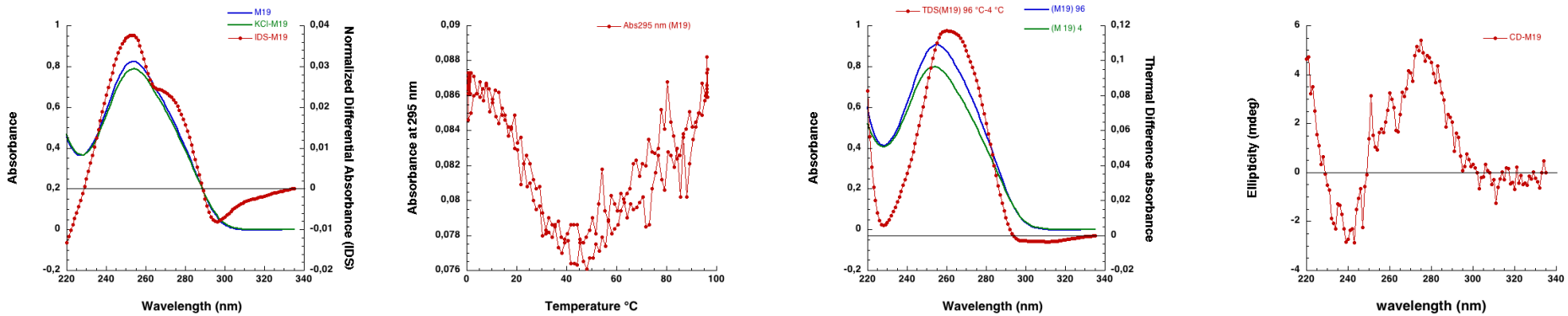
In vitro characterization of the selected candidates: **a)** Normalized Isothermal Difference Spectra (IDS) resulting from the difference between the absorbance recorded at 25°C before and after annealing in 100 mM KCl. **b)** Thermal melting transition (Tm) profiles measured at 240 and/or 295 nm (4 µM strand concentration, 10mM sodium cacodylate (pH= 7.0) and 100 mM KCl). **c)** Normalize differential spectra (TDS) in the 220 and 335 nm region (4 µM strand concentration, 10mM lithium cacodylate (pH= 7.0) and 100 mM KCl). **d)** Circular dichroism spectra (CD) (4 µM strand concentration, 10mM sodium cacodylate (pH= 7.0) and 100 mM KCl, 20°C). **e)** 1D Imino proton spectra in a 20 mM potassium phosphate buffer pH 6.9 with 70 mM KCl at 25°C. The presence of peacks between 10 and 12 ppm suggests G4 formation.

Table 26: Results interpretation of Marburg 18-1

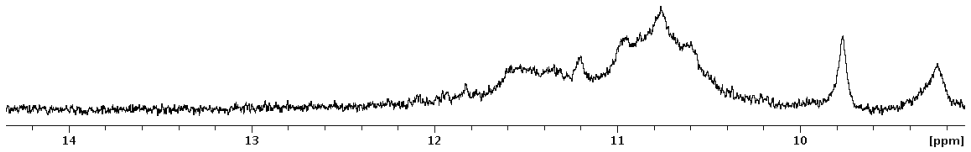
Technique	IDS	Tm	TDS	CD	NMR	**Conclusion**
Result (G4 ?)	Yes	Yes	Yes	Antiparallel	Yes	**G4**

Name: Marburg 19 **Sequence:** $^{5'}$ *TGAAGGGGAAGGAAGTGGTGCTCGGT* $^{3'}$ **Score:** 1.12

(a) (b) (c) (d)



(e)



In vitro characterization of the selected candidates: **a)** Normalized Isothermal Difference Spectra (IDS) resulting from the difference between the absorbance recorded at 25°C before and after annealing in 100 mM KCl. **b)** Thermal melting transition (Tm) profiles measured at 240 and/or 295 nm (4 µM strand concentration, 10mM sodium cacodylate (pH= 7.0) and 100 mM KCl). **c)** Normalize differential spectra (TDS) in the 220 and 335 nm region (4 µM strand concentration, 10mM lithium cacodylate (pH= 7.0) and 100 mM KCl). **d)** Circular dichroism spectra (CD) (4 µM strand concentration, 10mM sodium cacodylate (pH= 7.0) and 100 mM KCl, 20°C). **e)** 1D Imino proton spectra in a 20 mM potassium phosphate buffer pH 6.9 with 70 mM KCl at 25°C. The presence of peacks between 10 and 12 ppm suggests G4 formation.

Table 27: Results interpretation of Marburg 19

Technique	IDS	Tm	TDS	CD	NMR	**Conclusion**
Result (G4 ?)	Yes	Yes	No	No	Yes	**G4(Stable)**

Table 28: Add caption

Name	Start/End	Sequences	Length	Score	IDS	TM	TDS	CD	RMN	Conclusion
E1	2298-2323*	CGGTGGGGCGACAGTGGGTGTGCGG	25	1.32	Yes	40°C	Yes	No	Yes	G4
E1-1	2304-2331*	CGGGAGCCGGTGGGGCGACAGTGGG	25	1.36	No	No	No	Mixed	No	Not G4
E2	6980-7005	CGGGGAGTGGGCCTTCTGGGAA	22	1.32	No	No	No	No	Yes	Not G4
E3	7480-7509*	GTTTTGGGGACTTGTTGTGGTGGCGGGGT	29	1.41	Yes	35°C	Yes	Parallel	Yes	G4
E4	10646-10678	AGGGGTGGAAGGTTTATTGGGCTGGTATTG	30	1.23	Yes	30°C	Yes	Mixed	Yes	G4
E5	13901-13930	AGGGGTCATATGGGAGGGATTGAAGGA	27	1.41	Yes(-)	20°C	Yes(-)	Parallel	Yes	G4
M1	486-524	AGAGGGGGAGGATTGGGC	18	1.83	Yes(-)	Yes(-)	Yes(-)	Parallal	Yes	G4
M2	3423-3461*	CGCGGGTTGAGGAGGAGGGA	20	1.3	No	20°C	No	Parallel	Yes	G4
M3	4613-4653*	CGTGATCAGCATAAGGAGGAGGGTTCAAGT	30	0.57	No	No	No	No	No	Not G4
M4	6642-6679*	CGGATGGGCTGTGGGCAGTGGTAAAGGT	28	1.04	Yes	35°C	Yes	Antiparallel	Yes	G4
M5	6833-6870*	GCGTGCTTGGTTGTGGTGAGGGAGTGGGTGGC	32	1.03	Yes	35°C	Yes	Antiparallel	Yes	G4
M6	6938-6983*	AGAGATTGTAGTAGTGTTGTGAGGGGGCATGGACGGTTGTGCAGTTG	47	0.79	No	No	No	No	No	Not G4
M7	7083-7118*	TGTTGGAAGCAGGGTTGTTTTCGAGGGGGCACTGGT	36	1.03	No	No	No	No	No	Not G4
M8	7190-7242*	TGGGGGTGGGGGAGGGACTGGTGGA	25	2.24	Yes	55°c	Yes	Parallel	Not done	G4
M8-1	7194-7242*	CAAGATGTTGTGCAGTCGAGTTGGGGGTGGGGGAGGGACTGGTGGAATAC	50	1.18	Yes	50°C	No	Parallel	Yes	G4
M9	7848-7898	AGAGGGGACTGGTTGGGGTCTGGGTGGTAAATGGTGGA	38	1.47	Yes	30°C	Yes	Mixed	Yes	G4
M10	8962-8987*	ATGATTTGTGTAATGGGGTTTGGAGGG	27	1.19	No	10°C	No	No	Yes	?? G4
M11	9080-9119*	GCACATGTCCTTTGGGGGAGGAGGGGGTTTCGATAAGTTGA	41	1.02	No	No	No	No	No	Not G4
M12	10325-10360	AGTGTTGGGGGCTGGACAGTGAAGTGGGGA	30	1.43	Yes(-)	No	No	No	Yes	Not G4
M13	10809-10835	TGGGGTTTCTAGTGGAAGTCAGGAGGA	27	1.04	No	No	No	No	No	Not G4
M14	13361-13386	AGGGGAAAACGCTATAGTGAGAGGTGCGAG	30	0.8	No	No	No	No	No	Not G4
M15	15465-15498	TGGGTAAATACGCAGGGGGAGGTCAAGCTG	30	1.07	No	No	No	No	No	Not G4
M16	16865-16906*	TGGTTTGAGATGGGGAGGAGCCATGTATGTACGCAG	36	0.69	No	No	No	No	??	Not G4
M17	17339-17364	GCTTGGCCGAAGGAGAAGGAAGTGGTG	27	0.59	No	No	No	No	??	Not G4
M18	17341-17368	TTGGCCGAAGGGGAAGGAAGTGGTGCTCG	29	0.9	Yes	35°C	Yes	Antiparallel	Yes	Not G4
M18-1	17341-17375	TGGCTGAAGGGGAAGGAAGTGGTGCTCGGT	30	1.07	Yes	30°C	Yes (-)	Antiparallel	Yes	G4
M19	17346-17375	TGAAGGGGAAGGAAGTGGTGCTCGGT	26	1.12	Yes	20°C	No	No	Yes	G4

Bibliography

Bibliography

[1] J. D. Watson and F. H. Crick. Genetical implications of the structure of deoxyribonucleic acid. *Nature*, 171(4361):964–967, May 1953.

[2] R. E. Franklin and R. G. Gosling. Molecular configuration in sodium thymonucleate. *Nature*, 171(4356):740–741, Apr 1953.

[3] K. Hoogsteen. The structure of crystals containing a hydrogen-bounded complex of 1-methylthymine and 9-methyladenine. *Aceta Crystallographica*, 16:907–916, 1963.

[4] Vittorio Limongelli, Stefano De Tito, Linda Cerofolini, Marco Fragai, Bruno Pagano, Roberta Trotta, Sandro Cosconati, Luciana Marinelli, Ettore Novellino, Ivano Bertini, Antonio Randazzo, Claudio Luchinat, and Michele Parrinello. The g-triplex dna. *Angew Chem Int Ed Engl*, 52(8):2269–2273, Feb 2013.

[5] Maria Duca, Pierre Vekhoff, Kahina Oussedik, Ludovic Halby, and Paola B. Arimondo. The triple helix: 50 years later, the outcome. *Nucleic Acids Res*, 36(16):5123–5138, Sep 2008.

[6] Henry A. Day, Pavlos Pavlou, and Zoë A E. Waller. i-motif dna: structure, stability and targeting with ligands. *Bioorg Med Chem*, 22(16):4407–4418, Aug 2014.

[7] M. Gellert, M. N. Lipsett, and D. R. Davies. Helix formation by guanylic acid. *Proc Natl Acad Sci U S A*, 48:2013–2018, Dec 1962.

[8] Jeffery T. Davis. G-quartets 40 years later: from 5'-gmp to molecular biology and supramolecular chemistry. *Angew Chem Int Ed Engl*, 43(6):668–698, Jan 2004.

[9] D. Sen and W. Gilbert. Formation of parallel four-stranded complexes by guanine-rich motifs in dna and its implications for meiosis. *Nature*, 334(6180):364–366, Jul 1988.

[10] N. V. Hud, F. W. Smith, F. A. Anet, and J. Feigon. The selectivity for k+ versus na+ in dna quadruplexes is dominated by relative free energies of hydration: a thermodynamic analysis by 1h nmr. *Biochemistry*, 35(48):15383–15390, Dec 1996.

[11] C. C. Hardin, T. Watson, M. Corregan, and C. Bailey. Cation-dependent transition between the quadruplex and watson-crick hairpin forms of d(cgcg3gcg). *Biochemistry*, 31(3):833–841, Jan 1992.

[12] Alan Wong, Ramsey Ida, Lea Spindler, and Gang Wu. Disodium guanosine 5'-monophosphate self-associates into nanoscale cylinders at ph 8: a combined diffusion nmr spectroscopy and dynamic light scattering study. *J Am Chem Soc*, 127(19):6990–6998, May 2005.

[13] Vineeth Thachappilly Mukundan and Anh Tuan Phan. Bulges in g-quadruplexes: broadening the definition of g-quadruplex-forming sequences. *J Am Chem Soc*, 135(13):5017–5028, Apr 2013.

[14] Dinshaw J. Patel, Anh Tuân Phan, and Vitaly Kuryavyi. Human telomere, oncogenic promoter and 5'-utr g-quadruplexes: diverse higher order dna and rna targets for cancer therapeutics. *Nucleic Acids Res*, 35(22):7429–7455, 2007.

[15] Phillip A. Rachwal, I Stuart Findlow, Joern M. Werner, Tom Brown, and Keith R. Fox. Intramolecular dna quadruplexes with different arrangements of short and long loops. *Nucleic Acids Res*, 35(12):4214–4222, 2007.

[16] Aurore Guedin, Julien Gros, Patrizia Alberti, and Jean-Louis Mergny. How long is too long? effects of loop size on g-quadruplex stability. *Nucleic Acids Res*, 38(21):7858–7868, Nov 2010.

[17] Aurore Guedin, Patrizia Alberti, and Jean-Louis Mergny. Stability of intramolecular quadruplexes: sequence effects in the central loop. *Nucleic Acids Res*, 37(16):5559–5567, Sep 2009.

[18] Anh Tuân Phan. Human telomeric g-quadruplex: structures of dna and rna sequences. *FEBS J*, 277(5):1107–1117, Mar 2010.

[19] Julian L. Huppert and Shankar Balasubramanian. Prevalence of quadruplexes in the human genome. *Nucleic Acids Res*, 33(9):2908–2916, 2005.

[20] Anh Tuân Phan, Vitaly Kuryavyi, Hai Yan Gaw, and Dinshaw J. Patel. Small-molecule interaction with a five-guanine-tract g-quadruplex structure from the human myc promoter. *Nat Chem Biol*, 1(3):167–173, Aug 2005.

[21] Anh Tuan Phan, Vitaly Kuryavyi, Kim Ngoc Luu, and Dinshaw J. Patel. Structure of two intramolecular g-quadruplexes formed by natural human telomere sequences in k+ solution. *Nucleic Acids Res*, 35(19):6517–6525, 2007.

[22] Anh Tuan Phan, Vitaly Kuryavyi, Sarah Burge, Stephen Neidle, and Dinshaw J. Patel. Structure of an unprecedented g-quadruplex scaffold in the human c-kit promoter. *J Am Chem Soc*, 129(14):4386–4392, Apr 2007.

[23] C. Cheong and P. B. Moore. Solution structure of an unusually stable rna tetraplex containing g- and u-quartet structures. *Biochemistry*, 31(36):8406–8414, Sep 1992.

[24] Barbara Saccà, Laurent Lacroix, and Jean-Louis Mergny. The effect of chemical modifications on the thermal stability of different g-quadruplex-forming oligonucleotides. *Nucleic Acids Res*, 33(4):1182–1192, 2005.

[25] J. C. Darnell, K. B. Jensen, P. Jin, V. Brown, S. T. Warren, and R. B. Darnell. Fragile x mental retardation protein targets g quartet mrnas important for neuronal function. *Cell*, 107(4):489–499, Nov 2001.

[26] Oleg Kikin, Zachary Zappala, Lawrence D'Antonio, and Paramjeet S. Bagga. Grsdb2 and grsutrdb: databases of quadruplex forming g-rich sequences in pre-mrnas and mrnas. *Nucleic Acids Res*, 36(Database issue):D141–D148, Jan 2008.

[27] Herry Martadinata and Anh Tuan Phan. Formation of a stacked dimeric g-quadruplex containing bulges by the 5'-terminal region of human telomerase rna (hterc). *Biochemistry*, 53(10):1595–1600, Mar 2014.

[28] Brian Luke and Joachim Lingner. Terra: telomeric repeat-containing rna. *EMBO J*, 28(17):2503–2510, Sep 2009.

[29] Pooja Rawal, Veera Bhadra Rao Kummarasetti, Jinoy Ravindran, Nirmal Kumar, Kangkan Halder, Rakesh Sharma, Mitali Mukerji, Swapan Kumar Das, and Shantanu Chowdhury. Genome-wide prediction of g4 dna as regulatory motifs: role in escherichia coli global regulation. *Genome Res*, 16(5):644–655, May 2006.

[30] J. L. Huppert. Hunting g-quadruplexes. *Biochimie*, 90(8):1140–1148, Aug 2008.

[31] Titia de Lange. Shelterin: the protein complex that shapes and safeguards human telomeres. *Genes Dev*, 19(18):2100–2110, Sep 2005.

[32] R. J. Wellinger and D. Sen. The dna structures at the ends of eukaryotic chromosomes. *Eur J Cancer*, 33(5):735–749, Apr 1997.

[33] 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. *Proc Natl Acad Sci U S A*, 85(18):6622–6626, Sep 1988.

[34] M. G. Schechtman. Characterization of telomere dna from neurospora crassa. *Gene*, 88(2):159–165, Apr 1990.

[35] J. Forney, E. R. Henderson, and E. H. Blackburn. Identification of the telomeric sequence of the acellular slime molds didymium iridis and physarum polycephalum. *Nucleic Acids Res*, 15(22):9143–9152, Nov 1987.

[36] S. M. Le Blancq, R. S. Kase, and L. H. Van der Ploeg. Analysis of a giardia lamblia rrna encoding telomere with [taggg]n as the telomere repeat. *Nucleic Acids Res*, 19(20):5790, Oct 1991.

[37] E. J. Richards and F. M. Ausubel. Isolation of a higher eukaryotic telomere from arabidopsis thaliana. *Cell*, 53(1):127–136, Apr 1988.

[38] M. E. Petracek, P. A. Lefebvre, C. D. Silflow, and J. Berman. Chlamydomonas telomere sequences are a+t-rich but contain three consecutive g-c base pairs. *Proc Natl Acad Sci U S A*, 87(21):8222–8226, Nov 1990.

[39] C. Wicky, A. M. Villeneuve, N. Lauper, L. Codourey, H. Tobler, and F. Müller. Telomeric repeats (ttaggc)n are sufficient for chromosome capping function in caenorhabditis elegans. *Proc Natl Acad Sci U S A*, 93(17):8983–8988, Aug 1996.

[40] Neal F. Lue. Plasticity of telomere maintenance mechanisms in yeast. *Trends Biochem Sci*, 35(1):8–17, Jan 2010.

[41] Nancy Maizels and Lucas T. Gray. The g4 genome. *PLoS Genet*, 9(4):e1003468, Apr 2013.

[42] Jasmine S. Smith, Qijun Chen, Liliya A. Yatsunyk, John M. Nicoludis, Mark S. Garcia, Ramon Kranaster, Shankar Balasubramanian, David Monchaud, Marie-Paule Teulade-Fichou, Lara Abramowitz, David C. Schultz, and F Brad Johnson. Rudimentary g-quadruplex-based telomere capping in saccharomyces cerevisiae. *Nat Struct Mol Biol*, 18(4):478–485, Apr 2011.

[43] C. W. Greider and E. H. Blackburn. Identification of a specific telomere terminal transferase activity in tetrahymena extracts. *Cell*, 43(2 Pt 1):405–413, Dec 1985.

[44] N. W. Kim, M. A. Piatyszek, K. R. Prowse, C. B. Harley, M. D. West, P. L. Ho, G. M. Coviello, W. E. Wright, S. L. Weinrich, and J. W. Shay. Specific association of human telomerase activity with immortal cells and cancer. *Science*, 266(5193):2011–2015, Dec 1994.

[45] A. M. Zahler, J. R. Williamson, T. R. Cech, and D. M. Prescott. Inhibition of telomerase by g-quartet dna structures. *Nature*, 350(6320):718–720, Apr 1991.

[46] Kim Ngoc Luu, Anh Tuan Phan, Vitaly Kuryavyi, Laurent Lacroix, and Dinshaw J. Patel. Structure of the human telomere in k+ solution: an intramolecular (3 + 1) g-quadruplex scaffold. *J Am Chem Soc*, 128(30):9963–9970, Aug 2006.

[47] Tracy A. Brooks, Samantha Kendrick, and Laurence Hurley. Making sense of g-quadruplex and i-motif functions in oncogene promoters. *FEBS J*, 277(17):3459–3469, Sep 2010.

[48] Bruce A. Armitage. The rule of four. *Nat Chem Biol*, 3(4):203–204, Apr 2007.

[49] Julian L. Huppert and Shankar Balasubramanian. G-quadruplexes in promoters throughout the human genome. *Nucleic Acids Res*, 35(2):406–413, 2007.

[50] Adam Siddiqui-Jain, Cory L. Grand, David J. Bearss, and Laurence H. Hurley. 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(18):11593–11598, Sep 2002.

[51] Sarah Rankin, Anthony P. Reszka, Julian Huppert, Mire Zloh, Gary N. Parkinson, Alan K. Todd, Sylvain Ladame, Shankar Balasubramanian, and Stephen Neidle. Putative dna quadruplex formation within the human c-kit oncogene. *J Am Chem Soc*, 127(30):10584–10589, Aug 2005.

[52] Thomas S. Dexheimer, Daekyu Sun, and Laurence H. Hurley. Deconvoluting the structural and drug-recognition complexity of the g-quadruplex-forming region upstream of the bcl-2 p1 promoter. *J Am Chem Soc*, 128(16):5404–5415, Apr 2006.

[53] Daekyu Sun, Kexiao Guo, Jadrian J. Rusche, and Laurence H. Hurley. Facilitation of a structural transition in the polypurine/polypyrimidine tract within the proximal promoter region of the human vegf gene by the presence of potassium and g-quadruplex-interactive agents. *Nucleic Acids Res*, 33(18):6070–6080, 2005.

[54] Daekyu Sun, Kexiao Guo, and Yoon-Joo Shin. Evidence of the formation of g-quadruplex structures in the promoter region of the human vascular endothelial growth factor gene. *Nucleic Acids Res*, 39(4):1256–1265, Mar 2011.

[55] A. Lew, W. J. Rutter, and G. C. Kennedy. Unusual dna structure of the diabetes susceptibility locus iddm2 and its effect on transcription by the insulin promoter factor pur-1/maz. *Proc Natl Acad Sci U S A*, 97(23):12508–12512, Nov 2000.

[56] B. Westermark, A. Siegbahn, C. H. Heldin, and L. Claesson-Welsh. B-type receptor for platelet-derived growth factor mediates a chemotactic response by means of ligand-induced activation of the receptor protein-tyrosine kinase. *Proc Natl Acad Sci U S A*, 87(1):128–132, Jan 1990.

[57] Yuwei Chen, Prashansa Agrawal, Robert V. Brown, Emmanuel Hatzakis, Laurence Hurley, and Danzhou Yang. The major g-quadruplex formed in the human platelet-derived growth factor receptor ? promoter adopts a novel broken-strand structure in k+ solution. *J Am Chem Soc*, 134(32):13220–13223, Aug 2012.

[58] Erawan Borkham-Kamphorst, Jens Herrmann, Doris Stoll, Jens Treptau, Axel M. Gressner, and Ralf Weiskirchen. Dominant-negative soluble pdgf-beta receptor inhibits hepatic stellate cell activation and attenuates liver fibrosis. *Lab Invest*, 84(6):766–777, Jun 2004.

[59] G. Pesole, F. Mignone, C. Gissi, G. Grillo, F. Licciulli, and S. Liuni. Structural and functional features of eukaryotic mrna untranslated regions. *Gene*, 276(1-2):73–81, Oct 2001.

[60] Sunita Kumari, Anthony Bugaut, Julian L. Huppert, and Shankar Balasubramanian. An rna g-quadruplex in the 5' utr of the nras proto-oncogene modulates translation. *Nat Chem Biol*, 3(4):218–221, Apr 2007.

[61] P. Bois and A. J. Jeffreys. Minisatellite instability and germline mutation. *Cell Mol Life Sci*, 55(12):1636–1648, Sep 1999.

[62] Samir Amrane, Michael Adrian, Brahim Heddi, Alexandre Serero, Alain Nicolas, Jean-Louis Mergny, and Anh Tuan Phan. Formation of pearl-necklace monomorphic g-quadruplexes in the human ceb25 minisatellite. *J Am Chem Soc*, 134(13):5807–5816, Apr 2012.

[63] Anthony D. Keefe, Supriya Pai, and Andrew Ellington. Aptamers as therapeutics. *Nat Rev Drug Discov*, 9(7):537–550, Jul 2010.

[64] Guizhi Zhu, Mao Ye, Michael J. Donovan, Erqun Song, Zilong Zhao, and Weihong Tan. Nucleic acid aptamers: an emerging frontier in cancer therapy. *Chem Commun (Camb)*, 48(85):10472–10480, Nov 2012.

[65] L. C. Griffin, G. F. Tidmarsh, L. C. Bock, J. J. Toole, and L. L. Leung. In vivo anticoagulant properties of a novel nucleotide-based thrombin inhibitor and demonstration of regional anticoagulation in extracorporeal circuits. *Blood*, 81(12):3271–3276, Jun 1993.

[66] L. C. Bock, L. C. Griffin, J. A. Latham, E. H. Vermaas, and J. J. Toole. Selection of single-stranded dna molecules that bind and inhibit human thrombin. *Nature*, 355(6360):564–566, Feb 1992.

[67] Amandine Renaud de la Faverie. *Application de la technique du SELEX dans l'étude des quadruplexes de guanines*. PhD thesis, UNIVERSITÉ de BORDEAUX, 2013.

[68] Valeria Romanucci, Maria Gaglione, Anna Messere, Nicoletta Potenza, Armando Zarrelli, Sam Noppen, Sandra Liekens, Jan Balzarini, and Giovanni Di Fabio. Hairpin oligonucleotides forming g-quadruplexes: new aptamers with anti-hiv activity. *Eur J Med Chem*, 89:51–58, Jan 2015.

[69] Savvas N. Georgiades, Nurul H. Abd Karim, Kogularamanan Suntharalingam, and Ramon Vilar. Interaction of metal complexes with g-quadruplex dna. *Angew Chem Int Ed Engl*, 49(24):4020–4034, Jun 2010.

[70] Claudia Sissi, Barbara Gatto, and Manlio Palumbo. The evolving world of protein-g-quadruplex recognition: a medicinal chemist's perspective. *Biochimie*, 93(8):1219–1230, Aug 2011.

[71] P. Mohaghegh, J. K. Karow, RM Brosh, Jr, V. A. Bohr, and I. D. Hickson. The bloom's and werner's syndrome proteins are dna structure-specific helicases. *Nucleic Acids Res*, 29(13):2843–2849, Jul 2001.

[72] Cyril Ribeyre, Judith Lopes, Jean-Baptiste Boulé, Aurèle Piazza, Aurore Guédin, Virginia A. Zakian, Jean-Louis Mergny, and Alain Nicolas. The yeast pif1 helicase prevents genomic instability caused by g-quadruplex-forming ceb1 sequences in vivo. *PLoS Genet*, 5(5):e1000475, May 2009.

[73] Qian Li, Jun-Feng Xiang, Qian-Fan Yang, Hong-Xia Sun, Ai-Jiao Guan, and Ya-Lin Tang. G4ldb: a database for discovering and studying g-quadruplex ligands. *Nucleic Acids Res*, 41(Database issue):D1115–D1123, Jan 2013.

[74] Anne De Cian, Laurent Lacroix, Céline Douarre, Nassima Temime-Smaali, Chantal Trentesaux, Jean-François Riou, and Jean-Louis Mergny. Targeting telomeres and telomerase. *Biochimie*, 90(1):131–155, Jan 2008.

[75] Vijay Sekaran, Joana Soares, and Michael B. Jarstfer. Telomere maintenance as a target for drug discovery. *J Med Chem*, 57(3):521–538, Feb 2014.

[76] C. Schaffitzel, I. Berger, J. Postberg, J. Hanes, H. J. Lipps, and A. Plückthun. In vitro generated antibodies specific for telomeric guanine-quadruplex dna react with stylonychia lemnae macronuclei. *Proc Natl Acad Sci U S A*, 98(15):8572–8577, Jul 2001.

[77] Himesh Fernando, Raphaël Rodriguez, and Shankar Balasubramanian. Selective recognition of a dna g-quadruplex by an engineered antibody. *Biochemistry*, 47(36):9365–9371, Sep 2008.

[78] Stuart A. Hill and John K. Davies. Pilin gene variation in neisseria gonorrhoeae: reassessing the old paradigms. *FEMS Microbiol Rev*, 33(3):521–530, May 2009.

[79] John A. Capra, Katrin Paeschke, Mona Singh, and Virginia A. Zakian. G-quadruplex dna sequences are evolutionarily conserved and associated with distinct genomic features in saccharomyces cerevisiae. *PLoS Comput Biol*, 6(7):e1000861, 2010.

[80] Katrin Paeschke, John A. Capra, and Virginia A. Zakian. Dna replication through g-quadruplex motifs is promoted by the saccharomyces cerevisiae pif1 dna helicase. *Cell*, 145(5):678–691, May 2011.

[81] Raphaël Rodriguez, Kyle M. Miller, Josep V. Forment, Charles R. Bradshaw, Mehran Nikan, Sébastien Britton, Tobias Oelschlaegel, Blerta Xhemalce, Shankar Balasubramanian, and Stephen P. Jackson. Small-molecule-induced dna damage identifies alternative dna structures in human genes. *Nat Chem Biol*, 8(3):301–310, Mar 2012.

[82] Giulia Biffi, David Tannahill, John McCafferty, and Shankar Balasubramanian. Quantitative visualization of dna g-quadruplex structures in human cells. *Nat Chem*, 5(3):182–186, Mar 2013.

[83] Alexander Henderson, Yuliang Wu, Yu Chuan Huang, Elizabeth A. Chavez, Jesse Platt, F Brad Johnson, Robert M Brosh, Jr, Dipankar Sen, and Peter M. Lansdorp. Detection of g-quadruplex dna in mammalian cells. *Nucleic Acids Res*, 42(2):860–869, Jan 2014.

[84] Alan K. Todd. Bioinformatics approaches to quadruplex sequence location. *Methods*, 43(4):246–251, Dec 2007.

[85] Mathieu Fourment and Michael R. Gillings. A comparison of common programming languages used in bioinformatics. *BMC Bioinformatics*, 9:82, 2008.

[86] E. Quevillon, V. Silventoinen, S. Pillai, N. Harte, N. Mulder, R. Apweiler, and R. Lopez. Interproscan: protein domains identifier. *Nucleic Acids Res*, 33(Web Server issue):W116–W120, Jul 2005.

[87] R. Apweiler, T. K. Attwood, A. Bairoch, A. Bateman, E. Birney, M. Biswas, P. Bucher, L. Cerutti, F. Corpet, M. D. Croning, R. Durbin, L. Falquet, W. Fleischmann, J. Gouzy, H. Hermjakob, N. Hulo, I. Jonassen, D. Kahn, A. Kanapin, Y. Karavidopoulou, R. Lopez, B. Marx, N. J. Mulder, T. M. Oinn, M. Pagni, F. Servant, C. J. Sigrist, and E. M. Zdobnov. The interpro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res*, 29(1):37–40, Jan 2001.

[88] Alan K. Todd, Matthew Johnston, and Stephen Neidle. Highly prevalent putative quadruplex sequence motifs in human dna. *Nucleic Acids Res*, 33(9):2901–2907, 2005.

[89] Oleg Kikin, Lawrence D'Antonio, and Paramjeet S. Bagga. Qgrs mapper: a web-based server for predicting g-quadruplexes in nucleotide sequences. *Nucleic Acids Res*, 34(Web Server issue):W676–W682, Jul 2006.

[90] Camille Menendez, Scott Frees, and Paramjeet S. Bagga. Qgrs-h predictor: a web server for predicting homologous quadruplex forming g-rich sequence motifs in nucleotide sequences. *Nucleic Acids Res*, 40(Web Server issue):W96–W103, Jul 2012.

[91] Scott Frees, Camille Menendez, Matt Crum, and Paramjeet S. Bagga. Qgrs-conserve: a computational method for discovering evolutionarily conserved g-quadruplex motifs. *Hum Genomics*, 8(1):8, May 2014.

[92] Vinod Scaria, Manoj Hariharan, Amit Arora, and Souvik Maiti. Quadfinder: server for identification and analysis of quadruplex-forming motifs in nucleotide sequences. *Nucleic Acids Res*, 34(Web Server issue):W683–W685, Jul 2006.

[93] Anna Varizhuk, Dmitry Ischenko, Igor Smirnov, Olga Tatarinova, Vyacheslav Severov, Roman Novikov, Vladimir Tsvetkov, Vladimir Naumov, Dmitry Kaluzhny, and Galina Pozmogova. An improved search algorithm to find g-quadruplexes in genome sequences. *bioRxiv*, 2014.

[94] Johanna Eddy and Nancy Maizels. Gene function correlates with potential for g4 dna formation in the human genome. *Nucleic Acids Res*, 34(14):3887–3896, 2006.

[95] Steve G. Hershman, Qijun Chen, Julia Y. Lee, Marina L. Kozak, Peng Yue, Li-San Wang, and F Brad Johnson. Genomic distribution and functional analyses of potential g-quadruplex-forming sequences in saccharomyces cerevisiae. *Nucleic Acids Res*, 36(1):144–156, Jan 2008.

[96] Kajia Cao, Paul Ryvkin, and F. Brad Johnson. Computational detection and analysis of sequences with duplex-derived interstrand g-quadruplex forming potential. *Methods*, 57(1):3–10, May 2012.

[97] Jean-Denis Beaudoin, Rachel Jodoin, and Jean-Pierre Perreault. New scoring system to identify rna g-quadruplex folding. *Nucleic Acids Res*, 42(2):1209–1223, Jan 2014.

[98] Rumen Kostadinov, Nishtha Malhotra, Manuel Viotti, Robert Shine, Lawrence D'Antonio, and Paramjeet Bagga. Grsdb: a database of quadruplex forming g-rich sequences in alternatively processed mammalian pre-mrna sequences. *Nucleic Acids Res*, 34(Database issue):D119–D124, Jan 2006.

[99] Vinod Kumar Yadav, James Kappukalayil Abraham, Prithvi Mani, Rashi Kulshrestha, and Shantanu Chowdhury. Quadbase: genome-wide database of g4 dna–occurrence and conservation in human, chimpanzee, mouse and rat promoters and 146 microbes. *Nucleic Acids Res*, 36(Database issue):D381–D385, Jan 2008.

[100] Han Min Wong, Oliver Stegle, Simon Rodgers, and Julian Leon Huppert. A toolbox for predicting g-quadruplex formation and stability. *J Nucleic Acids*, 2010, 2010.

[101] Dengguo Wei, Alan K. Todd, Mire Zloh, Mekala Gunaratnam, Gary N. Parkinson, and Stephen Neidle. Crystal structure of a promoter sequence in the b-raf gene reveals an intertwined dimer quadruplex. *J Am Chem Soc*, 135(51):19319–19329, Dec 2013.

[102] Jean-Louis Mergny, Jing Li, Laurent Lacroix, Samir Amrane, and Jonathan B. Chaires. Thermal difference spectra: a specific signature for nucleic acid structures. *Nucleic Acids Res*, 33(16):e138, 2005.

[103] Jean-Louis Mergny and Laurent Lacroix. Analysis of thermal melting curves. *Oligonucleotides*, 13(6):515–537, 2003.

[104] Amandine Renaud de la Faverie, Aurore Guédin, Amina Bedrat, Liliya A. Yatsunyk, and Jean-Louis Mergny. Thioflavin t as a fluorescence light-up probe for g4 formation. *Nucleic Acids Res*, 42(8):e65, Apr 2014.

[105] Michael Adrian, Brahim Heddi, and Anh Tuan Phan. Nmr spectroscopy of g-quadruplexes. *Methods*, 57(1):11–24, May 2012.

[106] Jyotirmayee Mohanty, Nilotpal Barooah, V. Dhamodharan, S. Harikrishna, P. I. Pradeepkumar, and Achikanath C. Bhasikuttan. Thioflavin t as an efficient inducer and selective fluorescent sensor for the human telomeric g-quadruplex dna. *J Am Chem Soc*, 135(1):367–376, Jan 2013.

[107] Stefano Masiero, Roberta Trotta, Silvia Pieraccini, Stefano De Tito, Rosaria Perone, Antonio Randazzo, and Gian Piero Spada. A non-empirical chromophoric interpretation of cd spectra of dna g-quadruplex structures. *Org Biomol Chem*, 8(12):2683–2692, Jun 2010.

[108] Kah Wai Lim, Laurent Lacroix, Doris Jia En Yue, Joefina Kim Cheow Lim, Jocelyn Mei Wen Lim, and Anh Tuân Phan. Coexistence of two distinct g-quadruplex conformations in the htert promoter. *J Am Chem Soc*, 132(35):12331–12342, Sep 2010.

[109] Phong Lan Thao Tran, Antonella Virgilio, Veronica Esposito, Giuseppe Citarella, Jean-Louis Mergny, and Aldo Galeone. Effects of 8-methylguanine on structure, stability and kinetics of formation of tetramolecular quadruplexes. *Biochimie*, 93(3):399–408, Mar 2011.

[110] Jozef Nosek and Lubomír Tomáska. Mitochondrial genome diversity: evolution of the molecular architecture and replication strategy. *Curr Genet*, 44(2):73–84, Nov 2003.

[111] B. F. Lang, M. W. Gray, and G. Burger. Mitochondrial genome evolution and the origin of eukaryotes. *Annu Rev Genet*, 33:351–397, 1999.

[112] Gertraud Burger, Lise Forget, Yun Zhu, Michael W. Gray, and B Franz Lang. Unique mitochondrial genome architecture in unicellular relatives of animals. *Proc Natl Acad Sci U S A*, 100(3):892–897, Feb 2003.

[113] Gertraud Burger, Michael W. Gray, and B Franz Lang. Mitochondrial genomes: anything goes. *Trends Genet*, 19(12):709–716, Dec 2003.

[114] Justin C. St John, Joao Facucho-Oliveira, Yan Jiang, Richard Kelly, and Rana Salah. Mitochondrial dna transmission, replication and inheritance: a journey from the gamete through the embryo and into offspring and embryonic stem cells. *Hum Reprod Update*, 16(5):488–509, 2010.

[115] Dawei W. Dong, Filipe Pereira, Steven P. Barrett, Jill E. Kolesar, Kajia Cao, Joana Damas, Liliya A. Yatsunyk, F Brad Johnson, and Brett A. Kaufman. Association of g-quadruplex forming sequences with human mtdna deletion breakpoints. *BMC Genomics*, 15:677, 2014.

[116] Zhiyong Cheng and Michael Ristow. Mitochondria and metabolic homeostasis. *Antioxid Redox Signal*, 19(3):240–242, Jul 2013.

[117] Konstantinos Palikaras and Nektarios Tavernarakis. Mitochondrial homeostasis: The interplay between mitophagy and mitochondrial biogenesis. *Exp Gerontol*, Jan 2014.

[118] Siv G E. Andersson, Olof Karlberg, Bjorn Canback, and Charles G. Kurland. On the origin of mitochondria: a genomics perspective. *Philos Trans R Soc Lond B Biol Sci*, 358(1429):165–77; discussion 177–9, Jan 2003.

[119] S. Anderson, A. T. Bankier, B. G. Barrell, M. H. de Bruijn, A. R. Coulson, J. Drouin, I. C. Eperon, D. P. Nierlich, B. A. Roe, F. Sanger, P. H. Schreier, A. J. Smith, R. Staden, and I. G. Young. Sequence and organization of the human mitochondrial genome. *Nature*, 290(5806):457–465, Apr 1981.

[120] Robert W. Taylor and Doug M. Turnbull. Mitochondrial dna mutations in human disease. *Nat Rev Genet*, 6(5):389–402, May 2005.

[121] D. Bridge, C. W. Cunningham, B. Schierwater, R. DeSalle, and L. W. Buss. Class-level relationships in the phylum cnidaria: evidence from mitochondrial genome structure. *Proc Natl Acad Sci U S A*, 89(18):8750–8753, Sep 1992.

[122] Gertraud Burger and B Franz Lang. Parallels in genome evolution in mitochondria and bacterial symbionts. *IUBMB Life*, 55(4-5):205–212, 2003.

[123] J. C. Morris, M. E. Drew, M. M. Klingbeil, S. A. Motyka, T. T. Saxowsky, Z. Wang, and P. T. Englund. Replication of kinetoplast dna: an update for the new millennium. *Int J Parasitol*, 31(5-6):453–458, May 2001.

[124] T. M. Boyce, M. E. Zwick, and C. F. Aquadro. Mitochondrial dna in the bark weevils: size, structure and heteroplasmy. *Genetics*, 123(4):825–836, Dec 1989.

[125] J. E. Feagin. Mitochondrial genome diversity in parasites. *Int J Parasitol*, 30(4):371–390, Apr 2000.

[126] Daniel F. Bogenhagen, Denis Rousseau, and Stephanie Burke. The layered structure of human mitochondrial dna nucleoids. *J Biol Chem*, 283(6):3665–3675, Feb 2008.

[127] Takehiro Yasukawa, Ming-Yao Yang, Howard T. Jacobs, and Ian J. Holt. A bidirectional origin of replication maps to the major noncoding region of human mitochondrial dna. *Mol Cell*, 18(6):651–662, Jun 2005.

[128] Timothy A. Brown, Ciro Cecconi, Ariana N. Tkachuk, Carlos Bustamante, and David A. Clayton. Replication of mitochondrial dna occurs by strand displacement with alternative light-strand origins, not via a strand-coupled mechanism. *Genes Dev*, 19(20):2466–2476, Oct 2005.

[129] G. S. Shadel and D. A. Clayton. Mitochondrial dna maintenance in vertebrates. *Annu Rev Biochem*, 66:409–435, 1997.

[130] D. Ojala, C. Merkel, R. Gelfand, and G. Attardi. The trna genes punctuate the reading of genetic information in human mitochondrial dna. *Cell*, 22(2 Pt 2):393–403, Nov 1980.

[131] D. A. Clayton. Replication and transcription of vertebrate mitochondrial dna. *Annu Rev Cell Biol*, 7:453–478, 1991.

[132] Jessica Magnusson, Michael Orth, Patrick Lestienne, and Jan-Willem Taanman. Replication of mitochondrial dna occurs throughout the mitochondria of cultured human cells. *Exp Cell Res*, 289(1):133–142, Sep 2003.

[133] B. F. Lang, G. Burger, C. J. O'Kelly, R. Cedergren, G. B. Golding, C. Lemieux, D. Sankoff, M. Turmel, and M. W. Gray. An ancestral mitochondrial dna resembling a eubacterial genome in miniature. *Nature*, 387(6632):493–497, May 1997.

[134] J. L. Boore. Animal mitochondrial genomes. *Nucleic Acids Res*, 27(8):1767–1780, Apr 1999.

[135] Pedro H. Oliveira, Claudia Lobato da Silva, and Joaquim M S. Cabral. An appraisal of human mitochondrial dna instability: new insights into the role of non-canonical dna structures and sequence motifs. *PLoS One*, 8(3):e59907, 2013.

[136] G. L. Dianov, N. Souza-Pinto, S. G. Nyaga, T. Thybo, T. Stevnsner, and V. A. Bohr. Base excision repair in nuclear and mitochondrial dna. *Prog Nucleic Acid Res Mol Biol*, 68:285–297, 2001.

[137] M. Ingman, H. Kaessmann, S. Pääbo, and U. Gyllensten. Mitochondrial genome variation and the origin of modern humans. *Nature*, 408(6813):708–713, Dec 2000.

[138] Joana Damas, João Carneiro, Joana Gonçalves, James B. Stewart, David C. Samuels, António Amorim, and Filipe Pereira. Mitochondrial dna deletions are associated with non-b dna conformations. *Nucleic Acids Res*, 40(16):7606–7621, Sep 2012.

[139] Judith Lopes, Aurèle Piazza, Rodrigo Bermejo, Barry Kriegsman, Arianna Colosio, Marie-Paule Teulade-Fichou, Marco Foiani, and Alain Nicolas. G-quadruplex-induced instability during leading-strand replication. *EMBO J*, 30(19):4033–4046, Oct 2011.

[140] Subhajyoti De and Franziska Michor. Dna secondary structures and epigenetic determinants of cancer genome evolution. *Nat Struct Mol Biol*, 18(8):950–955, Aug 2011.

[141] David C. Samuels, Eric A. Schon, and Patrick F. Chinnery. Two direct repeats cause most human mtdna deletions. *Trends Genet*, 20(9):393–398, Sep 2004.

[142] Christophe Rocher, Thierry Letellier, William C. Copeland, and Patrick Lestienne. Base composition at mtdna boundaries suggests a dna triple helix model for human mitochondrial dna large-scale rearrangements. *Mol Genet Metab*, 76(2):123–132, Jun 2002.

[143] Paulina H. Wanrooij, Jay P. Uhler, Tomas Simonsson, Maria Falkenberg, and Claes M. Gustafsson. G-quadruplex structures in rna stimulate mitochondrial transcription termination and primer formation. *Proc Natl Acad Sci U S A*, 107(37):16072–16077, Sep 2010.

[144] Sanjay Kumar Bharti, Joshua A. Sommers, Jun Zhou, Daniel L. Kaplan, Johannes N. Spelbrink, Jean-Louis Mergny, and Robert 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*, 289(43):29975–29993, Oct 2014.

[145] Tobias Sing, Oliver Sander, Niko Beerenwinkel, and Thomas Lengauer. Rocr: visualizing classifier performance in r. *Bioinformatics*, 21(20):3940–3941, Oct 2005.

[146] Y. Huang and M. S. Pepe. A parametric roc model-based approach for evaluating the predictiveness of continuous markers in case-control studies. *Biometrics*, 65(4):1133–1144, Dec 2009.

[147] J. A. Swets, D. M. Green, D. J. Getty, and J. B. Swets. Signal detection and identification at successive stages of observation. *Percept Psychophys*, 23(4):275–289, Apr 1978.

[148] J. A. Swets. The relative operating characteristic in psychology: A technique for isolating effects of response bias finds wide use in the study of perception and cognition. *Science*, 182(4116):990–1000, Dec 1973.

[149] Paolo Sonego, András Kocsor, and Sándor Pongor. Roc analysis: applications to the classification of biological sequences and 3d structures. *Brief Bioinform*, 9(3):198–209, May 2008.

[150] Curk T. Vuk M. Roc curve, lift chart and calibration plot. *??*, 3:89–108, 2006.

[151] Daniel Berrar and Peter Flach. Caveats and pitfalls of roc analysis in clinical microarray research (and how to avoid them). *Brief Bioinform*, 13(1):83–97, Jan 2012.

[152] David Faraggi and Benjamin Reiser. Estimation of the area under the roc curve. *Stat Med*, 21(20):3093–3106, Oct 2002.

[153] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 44(3):837–845, Sep 1988.

[154] Lynne M. Harris and Catherine J. Merrick. G-quadruplexes in pathogens: a common route to virulence control? *PLoS Pathog*, 11(2):e1004562, Feb 2015.

[155] Vitaly Kuryavyi, Laty A. Cahoon, H Steven Seifert, and Dinshaw J. Patel. Reca-binding pile g4 sequence essential for pilin antigenic variation forms monomeric and 5' end-stacked dimeric parallel g-quadruplexes. *Structure*, 20(12):2090–2102, Dec 2012.

[156] Rupali Walia and George Chaconas. Suggested role for g4 dna in recombinational switching at the antigenic variation locus of the lyme disease spirochete. *PLoS One*, 8(2):e57792, 2013.

[157] Nicolas Smargiasso, Valérie Gabelica, Christian Damblon, Frédéric Rosu, Edwin De Pauw, Marie-Paule Teulade-Fichou, J Alexandra Rowe, and Antoine Claessens. Putative dna g-quadruplex formation within the promoters of plasmodium falciparum var genes. *BMC Genomics*, 10:362, 2009.

[158] Julie Norseen, F Brad Johnson, and Paul M. Lieberman. Role for g-quadruplex rna binding by epstein-barr virus nuclear antigen 1 in dna replication and metaphase chromosome attachment. *J Virol*, 83(20):10336–10346, Oct 2009.

[159] Jinzhi Tan, Clemens Vonrhein, Oliver S. Smart, Gerard Bricogne, Michela Bollati, Yuri Kusov, Guido Hansen, Jeroen R. Mesters, Christian L. Schmidt, and Rolf Hilgenfeld. The sars-unique domain (sud) of sars coronavirus contains two macrodomains that bind g-quadruplexes. *PLoS Pathog*, 5(5):e1000428, May 2009.

[160] Regis A. Vilchez and Janet S. Butel. Emergent human pathogen simian virus 40 and its role in cancer. *Clin Microbiol Rev*, 17(3):495–508, table of contents, Jul 2004.

[161] Jason Plyler, Karl Jasheway, Bodin Tuesuwan, Jessica Karr, Jarryd S. Brennan, Sean M. Kerwin, and Wendi M. David. Real-time investigation of sv40 large t-antigen helicase activity using surface plasmon resonance. *Cell Biochem Biophys*, 53(1):43–52, 2009.

[162] Sara Artusi, Matteo Nadai, Rosalba Perrone, Maria Angela Biasolo, Giorgio Palù, Louis Flamand, Arianna Calistri, and Sara 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*, 118:123–131, Apr 2015.

[163] Anna Avino, Carme Fabrega, Maria Tintore, and Ramon Eritja. Thrombin binding aptamer, more than a simple aptamer: chemically modified derivatives and biomedical applications. *Curr Pharm Des*, 18(14):2036–2047, 2012.

[164] Allicia C. Girvan, Yun Teng, Lavona K. Casson, Shelia D. Thomas, Simone Jüliger, Mark W. Ball, Jon B. Klein, William M Pierce, Jr, Shirish S. Barve, and Paula J. Bates. Agro100 inhibits activation of nuclear factor-kappab (nf-kappab) by forming a complex with nf-kappab essential modulator (nemo) and nucleolin. *Mol Cancer Ther*, 5(7):1790–1799, Jul 2006.

[165] Bärbel S. Blaum, Winfried Wünsche, Andrew J. Benie, Yuri Kusov, Hannelore Peters, Verena Gauss-Müller, Thomas Peters, and Georg Sczakiel. Functional binding of hexanucleotides to 3c protease of hepatitis a virus. *Nucleic Acids Res*, 40(7):3042–3055, Apr 2012.

[166] Hye-Min Woo, Ki-Sun Kim, Jin-Moo Lee, Hee-Sup Shim, Seong-Je Cho, Won-Kyu Lee, Hyuk Wan Ko, Young-Sam Keum, Soo-Youl Kim, Prabuddha Pathinayake, Chul-Joong Kim, and Yong-Joo Jeong. Single-stranded dna aptamer that specifically binds to the influenza virus ns1 protein suppresses interferon antagonism. *Antiviral Res*, 100(2):337–345, Nov 2013.

[167] Hye-Min Woo, Jin-Moo Lee, Sanggyu Yim, and Yong-Joo Jeong. Isolation of single-stranded dna aptamers that distinguish influenza virus hemagglutinin subtype h1 from h5. *PLoS One*, 10(4):e0125060, 2015.

[168] M. L. Andreola, F. Pileur, C. Calmels, M. Ventura, L. Tarrago-Litvak, J. J. Toulmé, and S. Litvak. Dna aptamers selected against the hiv-1 rnase h display in vitro antiviral activity. *Biochemistry*, 40(34):10087–10094, Aug 2001.

[169] D. J. Schneider, J. Feigon, Z. Hostomsky, and L. Gold. High-affinity ssdna inhibitors of the reverse transcriptase of type 1 human immunodeficiency virus. *Biochemistry*, 34(29):9599–9610, Jul 1995.

[170] Daniel Michalowski, Rebecca Chitima-Matsiga, Daniel M. Held, and Donald H. Burke. Novel bimodular dna aptamers with guanosine quadruplexes inhibit phylogenetically diverse hiv-1 reverse transcriptases. *Nucleic Acids Res*, 36(22):7124–7135, Dec 2008.

[171] Thomas M A. Gronewold, Antje Baumgartner, Jessica Hierer, Saleta Sierra, Michael Blind, Frank Schäfer, Julia Blümer, Tina Tillmann, Anne Kiwitz, Rolf Kaiser, Martin Zabe-Kühn, Eckhard Quandt, and Michael Famulok. Kinetic binding analysis of aptamers targeting hiv-1 proteins by a combination of a microbalance array and mass spectrometry (mams). *J Proteome Res*, 8(7):3568–3577, Jul 2009.

[172] W. Xu and A. D. Ellington. Anti-peptide aptamers recognize amino acid sequence and bind a protein epitope. *Proc Natl Acad Sci U S A*, 93(15):7475–7480, Jul 1996.

[173] Anh Tuan Phan, Vitaly Kuryavyi, Jin-Biao Ma, Aurelie Faure, Marie-Line Andreola, and Dinshaw J. Patel. An interlocked dimeric parallel-stranded dna quadruplex: a potent inhibitor of hiv-1 integrase. *Proc Natl Acad Sci U S A*, 102(3):634–639, Jan 2005.

[174] Aurélie Faure-Perraud, Mathieu Métifiot, Sandrine Reigadas, Patricia Recordon-Pinson, Vincent Parissi, Michel Ventura, and Marie-Line Andréola. The guanine-quadruplex aptamer 93del inhibits hiv-1 replication ex vivo by interfering with viral entry, reverse transcription and integration. *Antivir Ther*, 16(3):383–394, 2011.

[175] Joseph M. Watts, Kristen K. Dang, Robert J. Gorelick, Christopher W. Leonard, Julian W Bess, Jr, Ronald Swanstrom, Christina L. Burch, and Kevin M. Weeks. Architecture and secondary structure of an entire hiv-1 rna genome. *Nature*, 460(7256):711–716, Aug 2009.

[176] Andrew Rambaut, David Posada, Keith A. Crandall, and Edward C. Holmes. The causes and consequences of hiv evolution. *Nat Rev Genet*, 5(1):52–61, Jan 2004.

[177] Viviana Simon, David D. Ho, and Quarraisha Abdool Karim. Hiv/aids epidemiology, pathogenesis, prevention, and treatment. *Lancet*, 368(9534):489–504, Aug 2006.

[178] Mathieu Métifiot, Christophe Marchand, and Yves Pommier. Hiv integrase inhibitors: 20-year landmark and challenges. *Adv Pharmacol*, 67:75–105, 2013.

[179] Mathieu Métifiot, Samir Amrane, Simon Litvak, and Marie-Line Andreola. G-quadruplexes in viruses: function and potential therapeutic applications. *Nucleic Acids Res*, 42(20):12352–12366, Nov 2014.

[180] Surendran Mahalingam, Jayesh Meanger, Paul S. Foster, and Brett A. Lidbury. The viral manipulation of the host cellular and immune environments to enhance propagation and survival: a focus on rna viruses. *J Leukoc Biol*, 72(3):429–439, Sep 2002.

[181] Suman Ranjan Das and Shahid Jameel. Biology of the hiv nef protein. *Indian J Med Res*, 121(4):315–332, Apr 2005.

[182] Joseph Nkeze, Lin Li, Zsigmond Benko, Ge Li, and Richard Y. Zhao. Molecular characterization of hiv-1 genome in fission yeast schizosaccharomyces pombe. *Cell Biosci*, 5:47, 2015.

[183] Jamal Tazi, Nadia Bakkour, Virginie Marchand, Lilia Ayadi, Amina Aboufirassi, and Christiane Branlant. Alternative splicing: regulation of hiv-1 multiplication as a target for therapeutic action. *FEBS J*, 277(4):867–876, Feb 2010.

[184] W. I. Sundquist and S. Heaphy. Evidence for interstrand quadruplex formation in the dimerization of human immunodeficiency virus 1 genomic rna. *Proc Natl Acad Sci U S A*, 90(8):3393–3397, Apr 1993.

[185] Wen Shen, Robert J. Gorelick, and Robert A. Bambara. Hiv-1 nucleocapsid protein increases strand transfer recombination by promoting dimeric g-quartet formation. *J Biol Chem*, 286(34):29838–29847, Aug 2011.

[186] Dorota Piekna-Przybylska, Gaurav Sharma, and Robert A. Bambara. Mechanism of hiv-1 rna dimerization in the central region of the genome and significance for viral evolution. *J Biol Chem*, 288(33):24140–24150, Aug 2013.

[187] Rosalba Perrone, Matteo Nadai, Jerrod A. Poe, Ilaria Frasson, Manlio Palumbo, Giorgio Palù, Thomas E. Smithgall, and Sara N. Richter. Formation of a unique cluster of g-quadruplex structures in the hiv-1 nef coding region: implications for antiviral activity. *PLoS One*, 8(8):e73121, 2013.

[188] Rosalba Perrone, Elena Butovskaya, Dirk Daelemans, Giorgio Palù, Christophe Pannecouque, and Sara N. Richter. Anti-hiv-1 activity of the g-quadruplex ligand braco-19. *J Antimicrob Chemother*, 69(12):3248–3258, Dec 2014.

[189] Samir Amrane, Abdelaziz Kerkour, Amina Bedrat, Brune Vialet, Marie-Line Andreola, and Jean-Louis 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*, 136(14):5249–5252, Apr 2014.

[190] Dorota Piekna-Przybylska, Mark A. Sullivan, Gaurav Sharma, and Robert A. Bambara. U3 region in the hiv-1 genome adopts a g-quadruplex structure in its rna and dna sequence. *Biochemistry*, 53(16):2581–2593, Apr 2014.

[191] Wen Shen, Robert J. Gorelick, and Robert A. Bambara. Hiv-1 nucleocapsid protein increases strand transfer recombination by promoting dimeric g-quartet formation. *J Biol Chem*, 286(34):29838–29847, Aug 2011.

[192] R. Marquet, J. C. Paillart, E. Skripkin, C. Ehresmann, and B. Ehresmann. Dimerization of human immunodeficiency virus type 1 rna involves sequences located upstream of the splice donor site. *Nucleic Acids Res*, 22(2):145–151, Jan 1994.

[193] Michael Bukrinsky. A hard way to the nucleus. *Mol Med*, 10(1-6):1–5, 2004.

[194] Sebastien Wurtzer, Armelle Goubard, Fabrizio Mammano, Sentob Saragosti, Denise Lecossier, Allan J. Hance, and François Clavel. Functional central polypurine tract provides downstream protection of the human immunodeficiency virus type 1 genome from editing by apobec3g and apobec3b. *J Virol*, 80(7):3679–3683, Apr 2006.

[195] Arivazhagan Rajendran, Masayuki Endo, Kumi Hidaka, Phong Lan Thao Tran, Jean-Louis Mergny, Robert J. Gorelick, and Hiroshi Sugiyama. Hiv-1 nucleocapsid proteins as molecular chaperones for tetramolecular antiparallel g-quadruplex formation. *J Am Chem Soc*, 135(49):18575–18585, Dec 2013.

[196] S. N. Richter, I. Frasson, and G. Palù. Strategies for inhibiting function of hiv-1 accessory proteins: a necessary route to aids therapy? *Curr Med Chem*, 16(3):267–286, 2009.

[197] Vandana Purohit Basu, Min Song, Lu Gao, Sean T. Rigby, Mark Nils Hanson, and Robert A. Bambara. Strand transfer events during hiv-1 reverse transcription. *Virus Res*, 134(1-2):19–38, Jun 2008.

[198] Rosalba Perrone, Matteo Nadai, Ilaria Frasson, Jerrod A. Poe, Elena Butovskaya, Thomas E. Smithgall, Manlio Palumbo, Giorgio Palù, and Sara N. Richter. A dynamic g-quadruplex region regulates the hiv-1 long terminal repeat promoter. *J Med Chem*, 56(16):6521–6530, Aug 2013.

[199] Kathryn Twigg Arrildt, Sarah Beth Joseph, and Ronald Swanstrom. The hiv-1 env protein: a coat of many colors. *Curr HIV/AIDS Rep*, 9(1):52–63, Mar 2012.

[200] Yudong Quan, Chen Liang, Bluma G. Brenner, and Mark A. Wainberg. Multidrug-resistant variants of hiv type 1 (hiv-1) can exist in cells as defective quasispecies and be rescued by superinfection with other defective hiv-1 variants. *J Infect Dis*, 200(9):1479–1483, Nov 2009.

[201] Yuxing Li, Stephen A. Migueles, Brent Welcher, Krisha Svehla, Adhuna Phogat, Mark K. Louder, Xueling Wu, George M. Shaw, Mark Connors, Richard T. Wyatt, and John R. Mascola. Broad hiv-1 neutralization mediated by cd4-binding site antibodies. *Nat Med*, 13(9):1032–1034, Sep 2007.

[202] L. Fan and K. Peden. Cell-free transmission of vif mutants of hiv-1. *Virology*, 190(1):19–29, Sep 1992.

[203] M. Tristem, A. Purvis, and D. L. Quicke. Complex evolutionary history of primate lentiviral vpr genes. *Virology*, 240(2):232–237, Jan 1998.

[204] M. Tristem, C. Marshall, A. Karpas, and F. Hill. Evolution of the primate lentiviruses: evidence from vpx and vpr. *EMBO J*, 11(9):3405–3412, Sep 1992.

[205] N. K. Heinzinger, M. I. Bukrinsky, S. A. Haggerty, A. M. Ragland, V. Kewalramani, M. A. Lee, H. E. Gendelman, L. Ratner, M. Stevenson, and M. Emerman. The vpr protein of human immunodeficiency virus type 1 influences nuclear localization of viral nucleic acids in nondividing host cells. *Proc Natl Acad Sci U S A*, 91(15):7311–7315, Jul 1994.

[206] J. He, S. Choe, R. Walker, P. Di Marzio, D. O. Morgan, and N. R. Landau. Human immunodeficiency virus type 1 viral protein r (vpr) arrests cells in the g2 phase of the cell cycle by inhibiting p34cdc2 activity. *J Virol*, 69(11):6705–6711, Nov 1995.

[207] F. Re, D. Braaten, E. K. Franke, and J. Luban. Human immunodeficiency virus type 1 vpr arrests the cell cycle in g2 by inhibiting the activation of p34cdc2-cyclin b. *J Virol*, 69(11):6859–6864, Nov 1995.

[208] S. A. Stewart, B. Poon, J. Y. Song, and I. S. Chen. Human immunodeficiency virus type 1 vpr induces apoptosis through caspase activation. *J Virol*, 74(7):3105–3111, Apr 2000.

[209] Richard Yuqi Zhao, Michael Bukrinsky, and Robert T. Elder. Hiv-1 viral protein r (vpr) & host cellular responses. *Indian J Med Res*, 121(4):270–286, Apr 2005.

[210] Kathleen C. Prins, Sebastien Delpeut, Daisy W. Leung, Olivier Reynard, Valentina A. Volchkova, St Patrick Reid, Parameshwaran Ramanan, Washington B. Cardenas, Gaya K. Amarasinghe, Viktor E. Volchkov, and Christopher F. Basler. Mutations abrogating vp35 interaction with double-stranded rna render ebola virus avirulent in guinea pigs. *J Virol*, 84(6):3004–3015, Mar 2010.

[211] Smita P. Soni, Emmanuel Adu-Gyamfi, Sylvia S. Yong, Clara S. Jee, and Robert V. Stahelin. The ebola virus matrix protein deeply penetrates the plasma membrane: an important step in viral egress. *Biophys J*, 104(9):1940–1949, May 2013.

[212] Emmanuel Adu-Gyamfi, Smita P. Soni, Clara S. Jee, Michelle A. Digman, Enrico Gratton, and Robert V. Stahelin. A loop region in the n-terminal domain of ebola virus vp40 is important in viral assembly, budding, and egress. *Viruses*, 6(10):3837–3854, Oct 2014.

[213] V. E. Volchkov. Processing of the ebola virus glycoprotein. *Curr Top Microbiol Immunol*, 235:35–47, 1999.

[214] P. E. Rollin, R. J. Williams, D. S. Bressler, S. Pearson, M. Cottingham, G. Pucak, A. Sanchez, S. G. Trappier, R. L. Peters, P. W. Greer, S. Zaki, T. Demarcus, K. Hendricks, M. Kelley, D. Simpson, T. W. Geisbert, P. B. Jahrling, C. J. Peters, and T. G. Ksiazek. Ebola (subtype reston) virus among quarantined nonhuman primates recently imported from the philippines to the united states. *J Infect Dis*, 179 Suppl 1:S108–S114, Feb 1999.

[215] Thomas W. Geisbert and Lisa E. Hensley. Ebola virus: new insights into disease aetiopathology and possible therapeutic interventions. *Expert Rev Mol Med*, 6(20):1–24, Sep 2004.

[216] M. A. Bwaka, M. J. Bonnet, P. Calain, R. Colebunders, A. De Roo, Y. Guimard, K. R. Katwiki, K. Kibadi, M. A. Kipasa, K. J. Kuvula, B. B. Mapanda, M. Massamba, K. D. Mupapa, J. J. Muyembe-Tamfum, E. Ndaberey, C. J. Peters, P. E. Rollin, E. Van den Enden, and E. Van den Enden. Ebola hemorrhagic fever in kikwit, democratic republic of the congo: clinical observations in 103 patients. *J Infect Dis*, 179 Suppl 1:S1–S7, Feb 1999.

[217] R. C. Baron, J. B. McCormick, and O. A. Zubeir. Ebola virus disease in southern sudan: hospital dissemination and intrafamilial spread. *Bull World Health Organ*, 61(6):997–1003, 1983.

[218] J. W. Huggins. Prospects for treatment of viral hemorrhagic fevers with ribavirin, a broad-spectrum antiviral drug. *Rev Infect Dis*, 11 Suppl 4:S750–S761, 1989.

[219] Sophie J. Smither, Lin S. Eastaugh, Jackie A. Steward, Michelle Nelson, Robert P. Lenk, and Mark S. Lever. Post-exposure efficacy of oral t-705 (favipiravir) against inhalational ebola virus infection in a mouse model. *Antiviral Res*, 104:153–155, Apr 2014.

[220] Travis K. Warren, Jay Wells, Rekha G. Panchal, Kelly S. Stuthman, Nicole L. Garza, Sean A. Van Tongeren, Lian Dong, Cary J. Retterer, Brett P. Eaton, Gianluca Pegoraro, Shelley Honnold, Shanta Bantia, Pravin Kotian, Xilin Chen, Brian R. Taubenheim, Lisa S. Welch, Dena M. Minning, Yarlagadda S. Babu, William P. Sheridan, and Sina Bavari. Protection against filovirus diseases by a novel broad-spectrum nucleoside analogue bcx4430. *Nature*, 508(7496):402–405, Apr 2014.

[221] Stephen K. Gire, Augustine Goba, Kristian G. Andersen, Rachel S G. Sealfon, Daniel J. Park, Lansana Kanneh, Simbirie Jalloh, Mambu Momoh, Mohamed Fullah, Gytis Dudas, Shirlee Wohl, Lina M. Moses, Nathan L. Yozwiak, Sarah Winnicki, Christian B. Matranga, Christine M. Malboeuf, James Qu, Adrianne D. Gladden, Stephen F. Schaffner, Xiao Yang, Pan-Pan Jiang, Mahan Nekoui, Andres Colubri, Moinya Ruth Coomber, Mbalu Fonnie, Alex Moigboi, Michael Gbakie, Fatima K. Kamara, Veronica Tucker, Edwin Konuwa, Sidiki Saffa, Josephine Sellu, Abdul Azziz Jalloh, Alice Kovoma, James Koninga, Ibrahim Mustapha, Kandeh Kargbo, Momoh Foday, Mohamed Yillah, Franklyn Kanneh, Willie Robert, James L B. Massally, Sinéad B. Chapman, James Bochicchio, Cheryl Murphy, Chad Nusbaum, Sarah Young, Bruce W. Birren, Donald S. Grant, John S. Scheiffelin, Eric S. Lander, Christian Happi, Sahr M. Gevao, Andreas Gnirke, Andrew Rambaut, Robert F. Garry, S Humarr Khan, and Pardis C. Sabeti. Genomic surveillance elucidates ebola virus origin and transmission during the 2014 outbreak. *Science*, 345(6202):1369–1372, Sep 2014.

[222] Phong Lan Thao Tran, Jean-Louis Mergny, and Patrizia Alberti. Stability of telomeric g-quadruplexes. *Nucleic Acids Res*, 39(8):3282–3294, Apr 2011.

[223] Oliver Stegle, Linda Payet, Jean-Louis Mergny, David J C. MacKay, and Julian Huppert Leon. Predicting and understanding the stability of g-quadruplexes. *Bioinformatics*, 25(12):i374–i382, Jun 2009.

[224] Attila Ambrus, Ding Chen, Jixun Dai, Roger A. Jones, and Danzhou Yang. Solution structure of the biologically relevant g-quadruplex element in the human c-myc promoter. implications for g-quadruplex stabilization. *Biochemistry*, 44(6):2048–2058, Feb 2005.

[225] Michael Adrian, Ding Jie Ang, Christopher J. Lech, Brahim Heddi, Alain Nicolas, and Anh Tuan Phan. Structure and conformational dynamics of a stacked dimeric g-quadruplex formed by the human ceb1 minisatellite. *J Am Chem Soc*, Apr 2014.

[226] Phong Lan Thao Tran, Eric Largy, Florian Hamon, Marie-Paule Teulade-Fichou, and Jean-Louis Mergny. Fluorescence intercalator displacement assay for screening g4 ligands towards a variety of g-quadruplex structures. *Biochimie*, 93(8):1288–1296, Aug 2011.

[227] Liyun Zhang, Jun Cheng Er, Krishna Kanta Ghosh, Wan Jun Chung, Jaeduk Yoo, Wang Xu, Wei Zhao, Anh Tuan Phan, and Young-Tae Chang. Discovery of a structural-element specific g-quadruplex "light-up" probe. *Sci Rep*, 4:3776, 2014.

[228] Vineeth Thachappilly Mukundan, Ngoc Quang Do, and Anh Tuan Phan. Hiv-1 integrase inhibitor t30177 forms a stacked dimeric g-quadruplex structure containing bulges. *Nucleic Acids Res*, 39(20):8984–8991, Nov 2011.

[229] Katarina Tluckova, Maja Marusic, Petra Tothova, Lubos Bauer, Primoz sket, Janez Plavec, and Viktor Viglasky. Human papillomavirus g-quadruplexes. *Biochemistry*, 52(41):7207–7216, Oct 2013.

[230] Jixun Dai, Thomas S. Dexheimer, Ding Chen, Megan Carver, Attila Ambrus, Roger A. Jones, and Danzhou Yang. An intramolecular g-quadruplex structure with mixed parallel/antiparallel g-strands formed in the human bcl-2 promoter region in solution. *J Am Chem Soc*, 128(4):1096–1098, Feb 2006.

[231] Vitaly Kuryavyi, Anh Tuan Phan, and Dinshaw J. Patel. Solution structures of all parallel-stranded monomeric and dimeric g-quadruplex scaffolds of the human c-kit2 promoter. *Nucleic Acids Res*, 38(19):6757–6773, Oct 2010.

[232] Daniel J-F Chinnapen Hyun-Wu Lee and Dipankar Sen. Structure–function investigation of a deoxyribozyme with dual chelatase and peroxidase activities. *Pure Appl. Chem.*, 76:1537–1545, 2004.

[233] Graham D. Balkwill, Thomas P. Garner, Huw E L. Williams, and Mark S. Searle. Folding topology of a bimolecular dna quadruplex containing a stable mini-hairpin motif within the diagonal loop. *J Mol Biol*, 385(5):1600–1615, Feb 2009.

[234] Eun-Ang Raiber, Ramon Kranaster, Enid Lam, Mehran Nikan, and Shankar Balasubramanian. A non-canonical dna structure is a binding motif for the transcription factor sp1 in vitro. *Nucleic Acids Res*, 40(4):1499–1508, Feb 2012.

[235] P. Schultze, R. F. Macaya, and J. Feigon. Three-dimensional solution structure of the thrombin-binding dna aptamer d(ggttggtgtggttgg). *J Mol Biol*, 235(5):1532–1547, Feb 1994.