G4-iM Grinder: When size and frequency matter. G-Quadruplex, i-Motif and higher order structure search and analysis tool

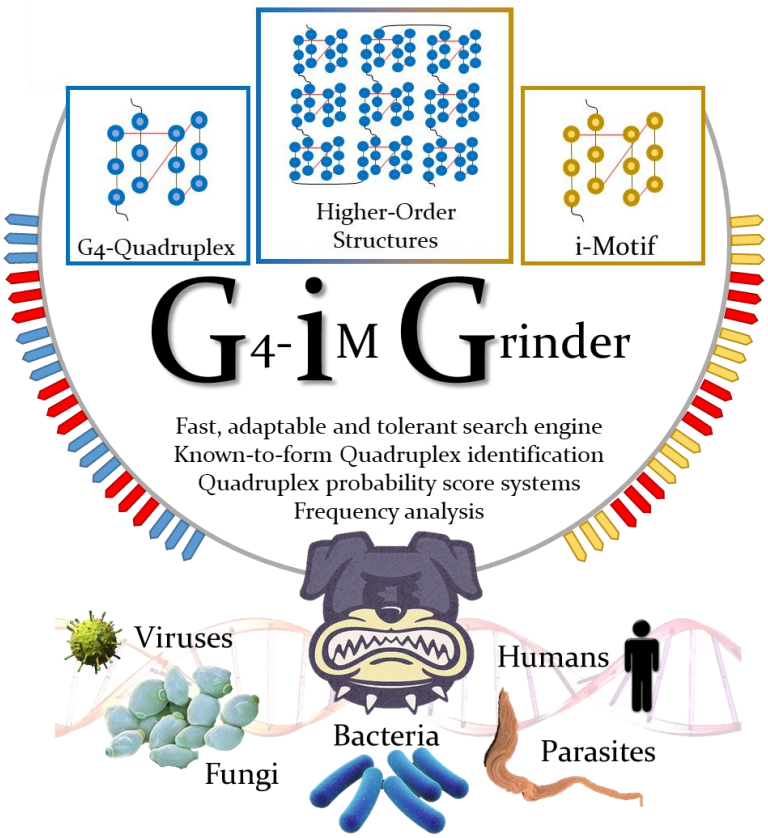
Efres Belmonte-Reche[1,2]* and Juan Carlos Morales[1].

[1] Department of Biochemistry and Molecular Pharmacology, Instituto de Parasitología y Biomedicina López Neyra, CSIC, PTS Granada, Avda. del Conocimiento, 17, 18016 Armilla, Granada, Spain.

[2] Life Sciences Department, International Iberian Nanotechnology Laboratory, Av. Mestre José Veiga, 4715-330 Braga, Portugal.

* To whom correspondence should be addressed; Phone, +351-253140112; e-mail: efres.belmonte@inl.int.

GRAPHICAL ABSTRACT



ABSTRACT

We present G4-iM Grinder, a system for the localization, characterization and selection of potential G4s, i-Motifs and higher-order structures. A robust and highly adaptable search engine identifies all structures that fit the user's quadruplex definitions. Their biological relevance, *in-vitro* formation probability and presence of known-to-form structures are then used as filters. The outcome is an efficient methodology that helps select the best candidates for a subsequent *in-vitro* analysis or a macroscopic genomic quadruplex assessment.

As proof of the analytical capabilities of G4-iM Grinder, the human genome was analysed for potential G4s and i-Motifs. Many known-to-form structures were identified. New candidates were selected considering their score and appearance frequency. We also focused on locating Potential Higher Order Quadruplex Sequences (PHOQS). We developed a new methodology to predict the most probable subunits of these assemblies and applied it to a PHOQS candidate.

Taking the human average density as reference, we examined the genomes of several etiological causes of disease. This first of its class comparative study found many organisms to be very dense in these potential quadruplexes. Many presented already known-to-form G4s and i-Motifs. These findings suggest the potential quadruplex have in the fight against these organisms that currently kill millions worldwide.

INTRODUCTION

Guanine rich nucleic acid sequences are capable of forming four-stranded structures called G-quadruplexes (G4), whilst cytosine based assemblies can form i-Motifs. These DNA and RNA conformations have been studied abundantly in the last few years due to the increasing evidence of their functional role in many living organisms (1, 2), yet the natural properties by which they form and work are very much unknown. To identify new structures, *in silico* predictions are based on *in vitro* verified paradigms (3–5). Loops (6), tetrad number, run imperfections (7) and the flanking regions of the structures (5, 8) all seem to play important roles in the topology and dynamics of these secondary structures.

Several tools for the identification of PQSs (putative G4 sequences) within a given DNA/RNA sequence are accessible to users nowadays. The first engines, such as Quadparser (9) and Quadfinder (10), were based on the folding rule which postulates that four perfect G-runs with shorter loops form the most stable G4s. Hence, results with these algorithms yield structures that usually fit the formula: (**G-run** {3:5} Loop {1:7})$_3$ **G-run** {3:5}, where the numbers inside the curly brackets are the range of acceptable lengths of the element.

However, many G4s have been identified that do not follow the folding rule. Loop range unconformity, G-run mismatches and bulges have been confirmed in several G4s (7), so a second generation of PQS search engines was designed to include them in the detection process.

QGRS Mapper (3, 11) partially addressed these irregularities by relaxing the folding rule to accept G-runs of size 2 and loop lengths of up to 45. The likelihood of G4 formation for each result is defined here through a scoring system that favours short and equal loop lengths and higher quartet presence. Similarly, Quadbase2 (12), ImGQfinder (13) and PQSfinder (4) also follow the folding rule (or a similar regular expression model). Of these, Quadbase2 and ImGQfinder are the more basic search engines that heavily restrict user-defined variable configuration. Quadbase2 can detect a fixed number of bulges within the G-runs of predefined size (3) following a regular expression model, and ImGQfinder considers both mismatches and bulges within G-runs in variating G-run sizes. PQSfinder, to the contrary, grants greater parameter liberty and at the same time tolerates G-run defects, such as bulges and mismatches, in the

detection process. Its scoring system has been proven to outmatch that of QGRS Mapper and is able to reduce false positive (PQS which are assumed to form G4 but do not) and false negative results (PQS which are assumed to be unable to form G4 but do). PQSfinder is also able to identify and resolve overlapping PQS, which is of utmost importance, as many G4 sequences overlap and compete for the common nucleotides to form the final structures (14).

Search engines that use the sliding window method and break with the folding rule have also been developed and used to detect potential G4s in a genome. Both G4 potential calculator (15) and G4Hunter (16) use this statistical analysis window that willingly defines neither individual PQS boundaries nor defect types. Hence, they can accommodate all G4-*errors* in the search at the expense of being unable to examine overlapping structures (as portions of nucleotides are analysed instead of regular sequences). Results found with G4 potential calculator are then analysed by their G-run density to determine G4-formation potential in a length independent manner. G4Hunter scoring system instead evaluates the result's G-richness and C-skewness to also consider the experimental destabilization effect caused by nearby cytosine presence on the G-quadruplex (as C can base pair with G and ultimately hinder G-quartet formation (17)).

The newest approach in the field is the development and use of G4-potential scoring methods based on machine-learning algorithms. These avoid predefined motif definitions and minimize formation assumptions to improve the analytical accuracy on non-standard PQSs, at the cost of obscurity in their predictive features. G4NN for example (18), employs an artificial neural network to classify the results of a sliding window model into forming and not-forming RNA G4 sequences. In a similar fashion, Quadron uses an artificial intelligence to classify folding rule abiding PQSs which returns the longest results for all the possible nested G4 sequences (5).

All quadruplex search models have several drawbacks and limitations despite the advances in the field. For the most part, variable configuration is usually heavily restricted meaning only the same kind of structures can be looked for (Table 1), excluding -for example- the detection of structures with more than four G-runs in the sequence. Even if only four G-runs

can form the G4-tetrads, extra G-runs can also occur in G-quadruplexes (19, 20) or as part of a fluctuating structure (21). Additionally, no current search engine considers nor calculates genomic PQS frequency from the results. Even if a higher frequency of a PQS does not mean a stronger tendency of *in vitro* G4 formation, it does mean that statistically they may be more biologically important or less biologically problematic. Also, higher G4 frequency allows easier and more accessible targets for the current G4-ligands, which in general are not selective between G4s (22–24). As example, we recently published the results of a PQS search in several parasitic genomes whilst considering frequency, and identified numerous highly recurrent potential G4 candidates (25). Most of these had already been described in literature as G4-forming structures; yet other sequences were new, including EBR1 which is repeated 33 times in the genome of *Trypanosoma brucei*. Despite EBR1 being graded poorly by the engine employed, we confirmed that the recurrent parasitic PQS was able to form G4 in solution even in the absence of cations. Other examples of the use of frequency as a selection filter also exist (26).

To conclude, none of the search engines have been explicitly designed to detect, analyse and evaluate higher order sequences. These assemblies with great biological potential are the result of very rich genomic G-tracks that form consecutive G4s. These can be assembled into a higher order structure formed by several G4 subunits. The human telomere sequence (hTel) higher order assembly is currently the main focus of this new area of investigation (27–30). Although several different models exist regarding the interactions between the units, the supra-structure has been found to influence the interactions between the hTel G4s and the telomeric proteins compared to individual G4s.

Table 1. Comparison of some of the search engines and analysers available for use. Structure qualification includes composition analysis and identification of sequences which are already known to form G4 *in vitro* within the results. Abbreviations: F. R. is folding rule, F. F. R. is flexible folding rule, S. W. is sliding window, A. N. N. is artificial neural network. Y stands for Yes, N for No.

	QGRS Mapper	G4Hunter	PQSfinder	G4RNA Screener	G4-iM Grinder
Format	web	R script/web	R package	Python script/web	R package
Search engine					
Model	F.R.	S.W.	F.F.R.	S.W.	F.F.R.
Run Composition	G	G, C	G	G	G, C, T, U, A
Run Imperfections	N	**Y**	**Y**	**Y**	**Y**
Modulable variables	5	2	10	3	**13**
Results analysis					
Structure analysis	**Y**	**Y**	**Y**	**Y**	**Y**
Structure frequency analysis	N	N	N	N	**Y**
High-order search & analysis	N	N	N	N	**Y**
Structure Qualification	N	N	N	N	**Y**
PQS Score	G-Score	G4Hunter	PQSfinder	G4NN (A. N. N.) G4Hunter cGcC	PQSfinder G4Hunter cGcC Frequency Total Score

MATERIAL AND METHODS

G4-iM Grinder

Our contribution to the field is focused on solving these limitations in an easy and fast manner for the user. G4-iM Grinder is an *in silico* tool designed as the starting step of the genomic quadruplex relevance study workflow. It is intended for researchers who want to detect quadruplex therapeutic targets (both known and new) on a genome, and efficiently filter and select the most interesting results to analyse *in vitro*. The final objective of the algorithm is to save time and resources in finding, evaluating, selecting and confirming these genomic structures.

Three distinct processes constitute G4-iM Grinder: the quadruplex search engine, the quadruplex qualification functions and the quantification functions. These processes tolerate parallelization over several cores to expedite the analysis.

Search engine: G4-iM Grinder's search engine was developed to allow extensive freedom in the user's definition of a quadruplex. These parameters are then applied to a fast, reliable and tolerant search motor capable of detecting even potential higher-order structures (Supplementary information – 1. G4-iM Grinder's search and analyser algorithm). The result is a very flexible algorithm capable of detecting all structures that fit the user's prerequisites, defined in 13 variables (Supplementary information – 2. Variables, predefined values and examples). These variables all have predefined values that conform to a functional yet broad definition of a flexible

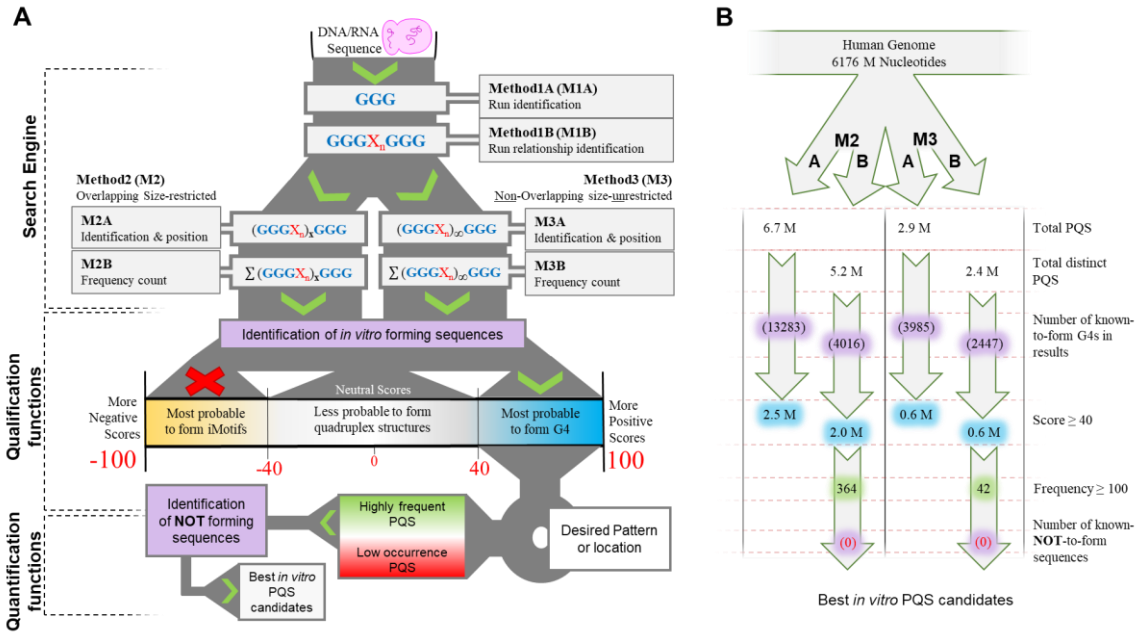
folding rule quadruplex disposition (Supplementary information – 3. G-Quadruplex and G4-iM Grinder). They can, however, be easily modified if, for example: structures with longer loops, more run bulges and/or shorter run sizes are also to be detected.

Method 1 (M1), Method 2 (M2) and Method 3 (M3) sub-processes constitute the search engine (Figure 1, A). M1 locates the runs (M1A) and finds their direct run-relationships (M1B), whilst M2 and M3 analyses the potential quadruplex structure formation. M2 does this analysis in an overlapping size-restricted manner (to detect quadruplex structures), and M3 in a non-overlapping size unrestricted way (to detect higher order sequences). In each case, the genomic location (M2A and M3A) and the genomic appearance frequency of each sequence is returned (M2B and M3B).

Quadruplex qualification and quantification functions: Several qualification functions were created, updated or adapted to narrow the number of interesting sequences after a search. The traditional qualification approach is to apply scoring systems that link high scores with high *in vitro* probability of formation as a selection or filtering mechanism. G4-iM Grinder incorporates some of these systems (G4hunter and cGcC) to evaluate its results. PQSfinder algorithm was upgraded using machine learning to overcome its current limitations and can also be used for this purpose. It is now capable of evaluating normal as well as irregular quadruplexes, i-Motifs and higher order sequences (Supplementary information – 5. Scoring models and their adaptations). A final score considering a weighted average formula of each scoring system and the frequency of the sequence (modulated by the variables *WeightedParameters* and *FreqWeight* respectively) can also be calculated as the sequence's quantitative interest score.

Other functions have been integrated into G4-iM Grinder to complement the selection process. The quantification of a predefined pattern as a percentage of the sequence (for example 'G', 'GGG' or 'TTA') and the localization of *in vitro* known-to-form quadruplex and known-**NOT**-to-form sequences can further help select the most interesting sequences (Supplementary information – 6. Explanation of other Variables).

Figure 1. **A.** G4-iM Grinder's workflow when *RunComposition* = G to find interesting PQS and PHOQS in a genome. Results are filtered by their scores, their frequency and the presence of known-to-form Quadruplex and known-NOT-to-form sequences. **B.** G4-iM Grinder's workflow results when applied to the human genome. Millions is abbreviated M.



Herein, we propose combining these G4-iM Grinder's capabilities in an efficient workflow. All sequences found within G4-iM Grinder's results with known-to-form quadruplex structures are already verified therapeutic targets ready for further investigation. Other sequences should be first filtered by their score. We found that a score of 40 or over is a good threshold that represents over 97 % of all known-to-form G4s examined (Supplementary information – 3. G-Quadruplex and G4-iM Grinder), and therefore represent the most plausible to form sequences. Genomic areas of interest (such as genes) or specific nucleotide arrangements within the quadruplex structure can be further used as filters. Alternatively, a frequency threshold can be employed that redirects the focus to the most recurrent high scoring structures. This way with few biophysical assays, many quadruplex targets with potentially greater biological repercussion can

be confirmed. Known-NOT-to-form sequences can then be matched to filter out sequences that coincide. The resulting sequences are therefore the most interesting results to start an quadruplex *in vitro* evaluation.

i-Motifs and G4-iM Grinder

C-rich regions of DNA or ARN have the ability to fold into tetrameric structures known as i-Motif (31, 32). These DNA assemblies consist of stranded duplexes sustained by hydrogen interactions between the intercalated nucleotide base pairs ***C***·***C***$^{+}$, which are stronger than the canonical G-C base pair when under acidic physiological conditions of temperature and ionic strength (33). Several consecutive ***C***s / ***C***$^{+}$s constitutes a run, and four **C-runs** arrange spatially as the final tetramer. The rest of the nucleotides in between **C-runs** form the minor or mayor grooves (loops). As an example, the telomeric sequence (**CCC**ATT)$_3$**CCC** (34), which can be organized as loops and runs, is able to form an i-Motif, yet the rich in C sequence CTCCCTTCTCCTCTC cannot (35).

G4-iM Grinder was designed to allow the search and evaluation of Putative i-Motifs Sequences (PiMS). Its search engine can locate all sequences that can form these quadruplexes using the same flexible folding rule used for G4s, as both are comprised of runs and loops. This follows previous uses of G4 search engines to detect PiMS for *in vitro* evaluation (36).

If the user wishes, G4-iM Grinder can use its qualification functions to evaluate the PiMS's *in vitro* formation potential. These include the application of the updated *PQSfinder*, *cGcC*, *G4Hunter* and *Final.Score* scoring algorithms, which operate (for i-Motifs) in an equal but contrary scale to G4s (where bigger negative values mean higher i-Motif probability; bigger positive values mean higher G4 probability).

Potentially, these scoring methods are useful for i-Motif punctuation as the algorithm's evaluation characteristics (designed for G4s) are also expected to influence C-based structure stability. For example, G4Hunter and cGcC analyse the sequence G and C relationships whilst PQSfinder

examines the tetrad size, bulges between tetrads and loop size to assess the potential of the sequences. 95 known-to-form i-Motifs published in literature were located, listed and analysed with G4-iM Grinder to test the use of these scoring systems in i-Motifs (Supplementary information – 4. i-Motifs and G4-iM Grinder). The mean score ± Standard Deviation (SD) of all these i-Motifs were: *G4Hunter* = -49.5 ± 17.0; *PQSfinder* = -62.1 ± 10.7; *score(mean)* = -55.8 ± 13.6, which indicates a strong relationship between the absolute high probability formation scores and the actual *in vitro* formation of the i-Motif (Supplementary information, Figure 4- Top: boxplots). These score relationships are very similar to that of the known-to-form G4s structures (Supplementary information – 3. Score (mean) = 52.9 ± 13.1) that validate the direct link between G4 *in vitro* formation and PQS score. In a similar fashion, these i-Motif results validate the link between i-Motif *in vitro* formation and the PiMS score. Furthermore, 87 % of i-Motifs scored over 40, and given the similar mean and SD to the G4s examined, we propose to use the same threshold of Score ≥ 40 to filter off these results. The pH dependent melting temperatures (Tm) of all available i-Motifs were also listed and used to find potential relationships with their scores and length. An evident direct correlation linking Tm and scores was measured (especially for pH 5, R^2 = 0.6), despite the deviating experimental conditions of the source's Tm determination (Supplementary information, Figure 4- bottom: graphs).

RESULTS

Full genomic analysis with G4-iM Grinder

An initial performance run was executed on human chromosome 22 to test the performance of the algorithm (Supplementary information – 7. G4-iM Grinder performance). Then, a full analysis of all the human chromosomes was carried out with predefined variable configuration for the identification of G-based PQS and C-based PiMS. Due to the small difference in results between using different *BulgeSize* values (as it is constrained by the variable

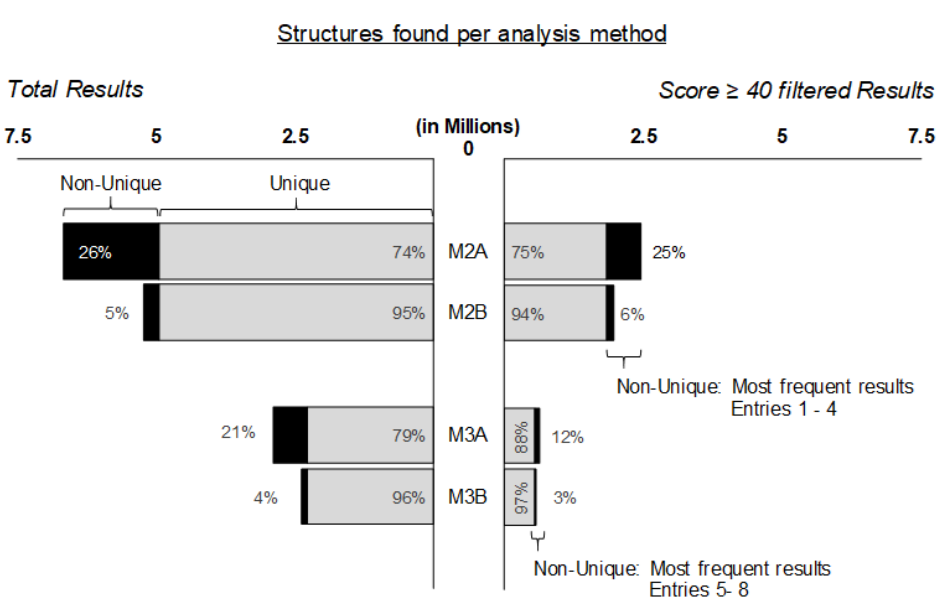
MaxIL, maximum number of bulges in a sequence), it was decided to accept only 1 different nucleotide per run, for a maximum total of 3 per sequence. It is well established that tetrad bulges are a factor for overall structure instability (37) and hence allowing too many of them would result in an increase in the detection rate of low probability *in vitro*-forming structures plus an increase in the computation processing time.

The complete human genomic analysis (M1, M2 and M3, A and B) took 12.6 h for G-based PQS and 10.4 h for C-based PiMS. In both cases, over 6.7 million potential sequences were detected with M2A (overlapping size-restricted method), and 2.9 million with M3A (non-overlapping size-unrestricted method - Figure 1, B). These results include thousands of confirmed G-quadruplexes and i-Motifs. Overall, the non-unique sequences (percentage of sequences with a frequency of occurrence of over 1) represent 26 and 21 % of all results (M2A and M3A respectively). Some of these sequences are repeated over 30000 times, although the average is 6.5 repetitions per non-unique sequence.

Following the workflow described previously, results were filtered by their scores to focus on the most probable to form *in vitro* sequences. Doing so for PQS and PiMS reduced the number of results found with M2 threefold whilst maintaining the ratio of unique to non-unique sequences. Of these (over 2 million results), 364 also had a frequency higher than 100, including the confirmed G-Quadruplexes T30695_or_T30923 (38), 20h (39), G4CT-pallidum (40), 93del (41) and 22Ag (42). None of these presented within known-NOT-to-form sequences. Similarly, applying the score filter decreased four fold (to 0.6 million) the original results of M3. Of these 0.6 million, only 42 sequences also had a frequency of at least 100. The most frequent sequences after applying these filters revealed interesting relationships, including the combination of entries 2 and 4 to give the bigger sequences of entries 6 and 7 of Figure 2, B. These combinations identify several widely spread high scoring sequences throughout the human genome, which can potentially form a changeable higher order structure or a fluctuating quadruplex.

Figure 2. **A.** Total found PQS (left) and found PQS after filtering by score (right) within the entire human genome per method of analysis. The results are divided into unique structures (found with a frequency of 1, in grey), and non-unique structures (found with a frequency of more than 1, in black). The percentage regarding the total is also shown in between parenthesis. **B.** Top four most frequent PQS found within the filtered by score results per method of analysis. Sequence portions in blue are the detected G-runs, in green the run bulges and in red the loops. Score is the average between G4Hunter and PQSfinder. Abbreviations: Freq is frequency, IL is total run bulges and %G is the percentage of G in the sequence.

A



B

Entry	Sequence	Freq	Score	Runs	IL	%G	Length
M2B *- Overlapping size restricted*							
1	GGGAGGGAGGGAGGG	13575	62	4	0	80	15
2	GGGGTGGGGGGAGGGGGGAGGG	6724	82	4	0	86	22
3	GGGAGGGAGGGAGGGAGGG	6446	62	5	0	79	19
4	GGGGACTGTTGTGGGGTGGGGGGAGGGGG	3691	72	4	0	72	29
M3B *- Non-overlapping size un-restricted*							
5	GGGAGGGAGGGAGGG	1084	62	4	0	80	15
6	GGGGACTGTTGTGGGGTGGGGGGAGGGGGGAGGGATAGCATTAGGAG	987	54	6	1	62	47
7	GGGGACTGTTGTGGGGTGGGGGGAGGGGGGAGGGATAGCATTGGG	955	62	6	0	64	45
8	GGGGTGGTGGCCGGGCAGAGGGG	952	50	4	1	70	23

The search for PiMS gave very similar results to that of PQS. Most noticeably, KRC6 (43), hTel (44), cJun (45) and C3T333 (46) were located within the 357 sequences found with M2 that score at least 40 and have a frequency of at least 100.

Potential Higher Order Quadruplex Sequences (PHOQS) and their analysis

Using M3A results, the longest of all possible higher order quadruplex sequences (PHOQS) was identified in chromosome 6. This structure potentially involves more than 2700 nucleotides and can be formed by over 300 possible PQS options, yet it was graded poorly because of its many bulges in between G-runs. Hence, the focus was set on the longest structures with the highest probability of formation (Score ≥ 50, Table 2). PHOQS found this way include a 2343 long sequence in chromosome 11 (entry 1) and a 1005 segment in the end of chromosome X, rich in the telomeric and other known-to-form G4 sequences (entry 10).

Table 2. Longest Potential Higher Order Quadruplex Sequence (PHOQS) found in the human genome, which scores at least 50 with method 3A (M3A). Conf.Quad.Seqs are identified G4 sequences within the found structure which are known to form *in vitro*, followed by the times detected in between parenthesis. Score is the mean of G4hunter and PQSfinder. Abbreviations: Chrom. is chromosome, IL is run bulges, %G, %C, %A and %T is the percentage of that nucleotide in the sequence.

Entry	Chrom.	Start	Length	Runs	IL	Strand	Score	%G	%C	%A	%T	Conf.Quad.Seqs
1	11	400747	2343	305	13	-	58	66	17	7	10	
2	X	1587610	1666	293	37	-	61	70	1	29	1	
3	20	64093767	1646	151	13	+	53	60	8	13	19	
4	2	240923448	1351	156	0	-	54	59	9	7	25	
5	X	328170	1240	182	16	-	68	74	5	21	1	
6	10	131041965	1170	148	2	+	56	55	2	22	21	
7	7	470172	1125	136	0	-	54	58	8	17	16	
8	8	141812709	1102	142	8	-	54	60	9	13	18	
9	9	133668264	1089	189	11	-	51	66	1	27	6	
10	X	156029890	1005	164	6	+	52	56	0	13	31	26gtel4 (4) 22Ag (68) Tet22 (7) Gia18 (2) Scer21 (1) 45Ag (51) 26gsc (1)
11	2	239737366	990	163	4	-	60	63	19	4	14	d(G4C2)4 (3)
12	3	10005	978	159	1	-	52	55	2	15	29	TSG24 (44) X3ACT (3) Tet22 (2) G4CT-pallidum (6)

Attention was set on HoEBR1, a relatively small sized (< 200 nucleotides to avoid excessive complexity), high scoring and frequent PHOQS. This 118 nucleotide-long PHOQS is repeated four times in the human complementary strand of chromosome 16. Here, it forms part of a nuclear pore complex interacting proteins (NPIPA1 and 2 genes), a polycystin 1 transient receptor potential channel and several other unidentified genes. HoEBR1 can be formed by a combination of its 32 potential PQS subunits (identified by extracting the results from M2A within the location of HoEBR1, Table 3). The known-to-form G4 sequence IV-1242540 was also located within these potential subunits (47).

Table 3. HoEBR1 analysis and dissection into its core possible PQS subunits. In black and in the first row HoEBR1, and beneath are all the possible PQS units that can potentially form the higher order structure. Sequence portions in blue are the detected G-runs, in green the run bulges and in red the loops. Score is the average between G4Hunter and PQSfinder. Abbreviations: C. Q. S. (confirmed quadruplex sequences) are identified G4 structures within the sequence which are known to form *in vitro*, followed by the times detected in between parenthesis, IL is total run bulges, %G, %C, %A and %T is the percentage of that nucleotide in the sequence.

Tag	Length	Runs	IL	Sequence	Score	%G	%C	%A	%T	C.Q.S.
HoEBR1	118	17	3	GGGTCTGGGGAAAGAAGAAGAGGAGGAGGAGGAGGGGTTGTCGGG GGAAGAGGAGGAAAGGGAAGGGAATGAAGGGGGGAAGGGGAGGGG AAGGGGAGGGGGAGGGGAGGGGGAGGGG	**62**	64	2	29	5	
Possible PQS subunits										
I	29	4	2	GGGTCTGGGGAAAGAAGAAGAGGAGGAGG	**30**	55	3	35	7	
II	31	4	2	GGGGAAAGAAGAAGAGGAGGAGGAGGAGGGG	**34**	61	0	39	0	
III	28	4	2	GAGGAGGAGGAGGAGGGGTTGTCGGGGG	**41**	68	4	18	11	
IV	26	4	2	GGAGGAGGAGGAGGGGTTGTCGGGGG	**46**	69	4	15	12	
V	32	5	3	GGAGGAGGAGGAGGGGTTGTCGGGGGAAGAGG	**38**	66	3	22	9	
VI	25	4	2	GAGGAGGAGGAGGGGTTGTCGGGGG	**44**	68	4	16	12	
VII	31	5	3	GAGGAGGAGGAGGGGTTGTCGGGGGAAGAGG	**36**	65	3	23	10	
VIII	29	4	2	GGAGGAGGAGGGGTTGTCGGGGGAAGAGG	**40**	66	3	21	10	
IX	28	4	2	GAGGAGGAGGGGTTGTCGGGGGAAGAGG	**38**	64	4	21	11	
X	26	4	2	GGAGGAGGGGTTGTCGGGGGAAGAGG	**42**	65	4	19	12	
XI	25	4	2	GAGGAGGGGTTGTCGGGGGAAGAGG	**40**	64	4	20	12	
XII	29	4	1	GGGGTTGTCGGGGGAAGAGGAGGAGGAAAGGG	**46**	62	3	24	10	
XIII	25	4	1	GGGGGAAGAGGAGGAAAGGGAAGGG	**46**	64	0	36	0	
XIV	29	4	1	GAGGAGGAAAGGGAAGGGAATGAAGGGGG	**42**	59	0	38	3	
XV	27	4	1	GGAGGAAAGGGAAGGGAATGAAGGGGG	**46**	59	0	37	4	
XVI	26	4	1	GAGGAAAGGGAAGGGAATGAAGGGGG	**44**	58	0	39	4	
XVII	33	5	1	GAGGAAAGGGAAGGGAATGAAGGGGGGAAGGGG	**49**	61	0	36	3	
XVIII	26	4	0	GGGAAGGGAATGAAGGGGGGAAGGGG	**62**	65	0	31	4	
XIX	31	5	0	GGGAAGGGAATGAAGGGGGGAAGGGGAGGGG	**64**	68	0	29	3	
XX	26	4	0	GGGAATGAAGGGGGGAAGGGGAGGGG	**67**	69	0	28	3	
XXI	32	5	0	GGGAATGAAGGGGGGAAGGGGAGGGGAAGGGG	**68**	69	0	27	4	
XXII	23	4	0	GGGGGGAAGGGGAGGGGAAGGGG	**78**	78	0	22	0	
XXIII	29	5	0	GGGGGGAAGGGGAGGGGAAGGGGAGGGGG	**80**	79	0	21	0	IV-1242540 (1)
XXIV	21	4	0	GGGGAGGGGAAGGGGAGGGGG	**80**	81	0	19	0	IV-1242540 (1)
XXV	26	5	0	GGGGAGGGGAAGGGGAGGGGGAGGGG	**80**	81	0	19	0	IV-1242540 (1)
XXVI	32	6	0	GGGGAGGGGAAGGGGAGGGGGAGGGGAGGGGG	**82**	81	0	19	0	IV-1242540 (1)
XXVII	21	4	0	GGGGAAGGGGAGGGGGAGGGG	**80**	81	0	19	0	
XXVIII	27	5	0	GGGGAAGGGGAGGGGGAGGGGAGGGGG	**82**	82	0	19	0	
XXIX	32	6	0	GGGGAAGGGGAGGGGGAGGGGAGGGGGAGGGG	**82**	81	0	19	0	
XXX	21	4	0	GGGGAGGGGGAGGGGAGGGGG	**86**	86	0	14	0	
XXXI	26	5	0	GGGGAGGGGGAGGGGAGGGGGAGGGG	**84**	85	0	15	0	
XXXII	21	4	0	GGGGGAGGGGAGGGGGAGGGG	**86**	86	0	14	0	

All these subunits overlap and will potentially compete to form the most stable structures. An algorithm was developed to predict the most interesting combinations of PQS subunits to form HoEBR1. Such a tool is included in the G4-iM Grinder package under the function *GiG.M3Structure*. The idea behind the code is to consider the PHOQS as several *seats* for which all the subunits are candidates. When a candidate claims a *seat*, it will annul any other candidate with which it shares nucleotides. In our case, HoEBR1 can be potentially be formed by up to four *seats* (Figure 3, A).

At first, *seat* allocation was decided to be sequential, assigning a *seat* first to the best scoring PQS with known-to-form G4 in their sequence (method HSA, Highest-score Sequential

Assignment). This process yielded a unique organizational candidate that presented three seats and a poor overall score due to election of subunit XXVI as first *seat*. This election ultimately hinders the formation of two other interesting subunits in the tail of HoEBR1 which lowers its overall score (Figure 3, B: 1. HSA Conformation).

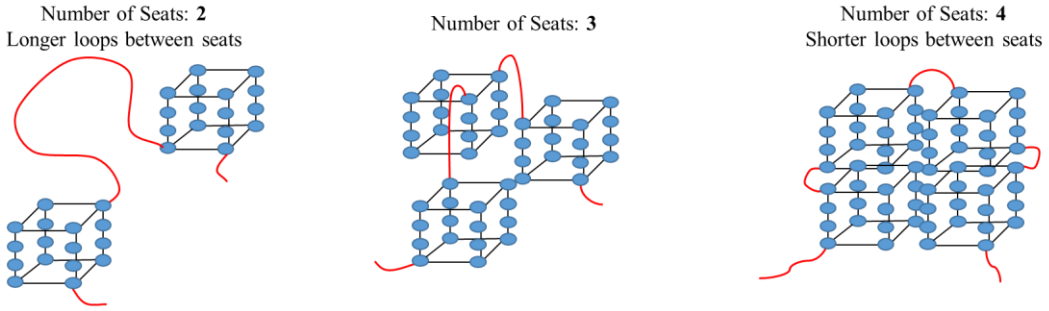
An alternative method based on randomly assigning *seats* to candidates was also developed and used. After 10000 iterations, the process identified all possible 307 subunit conformations that can give rise to HoEBR1. This was repeated ten times to make sure no conformations had been excluded. The 307 arrangements were then analysed by their mean *seat* PQS scores (Figure 3, B: Graph), as highest PQS scores are more probable to form G4 *in vitro* and therefore more probable to be the actual PHOQS subunits. Under such pretences, the highest mean score conformation is a three-*seat* structure composed by the PQSs: IV, XXII and XXXII (Figure 3, B: 2. RAH Conformation).

The RAH conformation is based solely on PQS scores and therefore does not consider the loop size between subunits in its study. It can be argued that (as happens within G4s) longer loops are likely to decrease overall stability of the greater structure. Hence, the scores of the conformations were also normalized by the percentage of the PHOQS which is involved as PQS for that given conformation (Figure 3, C: Graph). This way the method discriminates bigger loops between G4 in favour of higher PQS-density conformations. When applied to HoEBR1, 6 four-*seat* configurations scored highly ($Score_n > 50$, 2 % of total conformations; Figure 3, C), which are the results of electing 10 possible subunits (Figure 3, C: 3). The highest scoring (normalized) arrangement found this way was the combination of the PQS Candidates: I, XII, XXI and XXXII PQS (Figure 3, C: 4. RAnH Conformation). Here, over 96 % of its nucleotides are involved as PQSs and less than 3 % are loops between *seats*.

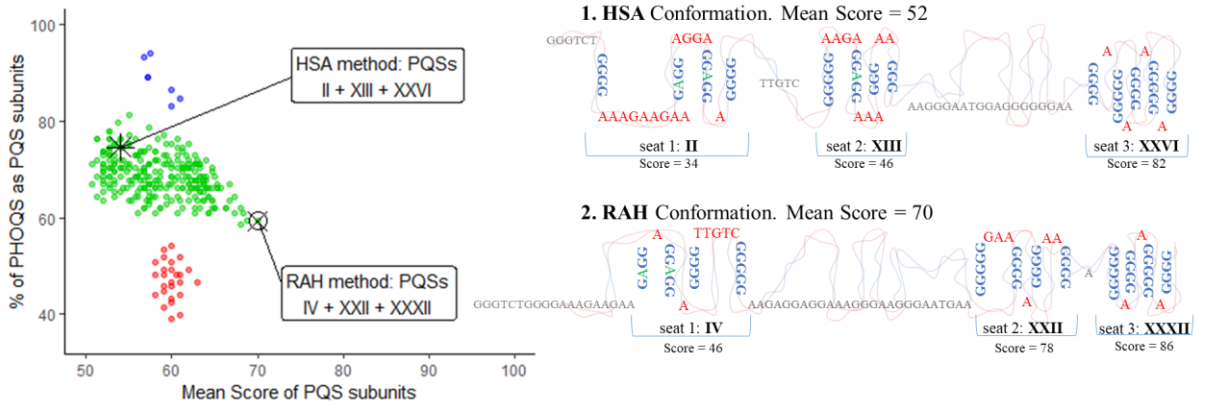
Figure 3. **A.** The PHOQS HoEBR1 can be arranged into up to four *seats*. Greater number of *seats* means smaller loops between units and potentially a gain in structure stability. **B. 1.** High-scoring Sequential Allocation (HSA) conformation is based on assigning *seats* sequentially to the highest

scoring candidate and known-to-form G4s. **Graph.** After 10000 iterations of random seat allocation, all 307 candidate conformations of HoEBR1 were found and studied by the mean PQS score of the candidates forming the conformation. In red, blue and green, conformations with 2, 3 and 4 *seats* respectively. **2.** Random Allocation High-scoring (RAH) conformation is the highest mean PQS scoring arrangement. **C. Graph.** The 307 conformations were normalized by the percentage of the PHOQS that is involved as a PQS to favour shorter loops between subunits and greater *seat* density. **3.** The focus was set on the best scoring conformations which present four *seats.* These can be occupied by a combination of 10 candidates. **4.** Random Allocation normalized High-scoring (RAnH) conformation is the mean normalized highest scoring arrangement. Topologies are not accurate.

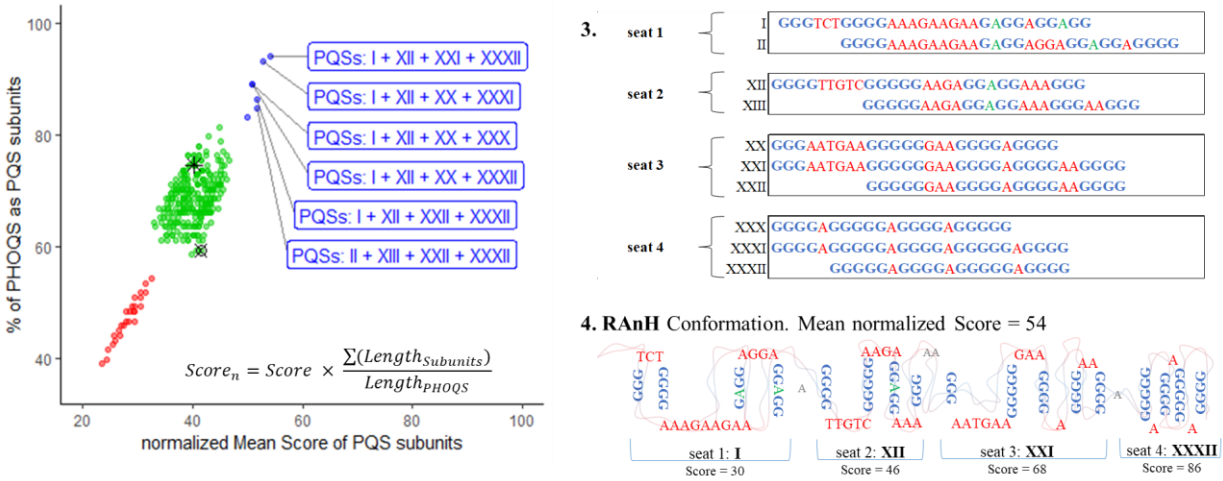
A.



B. Conformations of HoEBR1 arranged by their mean PQS score



C. Conformations of HoEBR1 arranged by their normalized mean PQS score



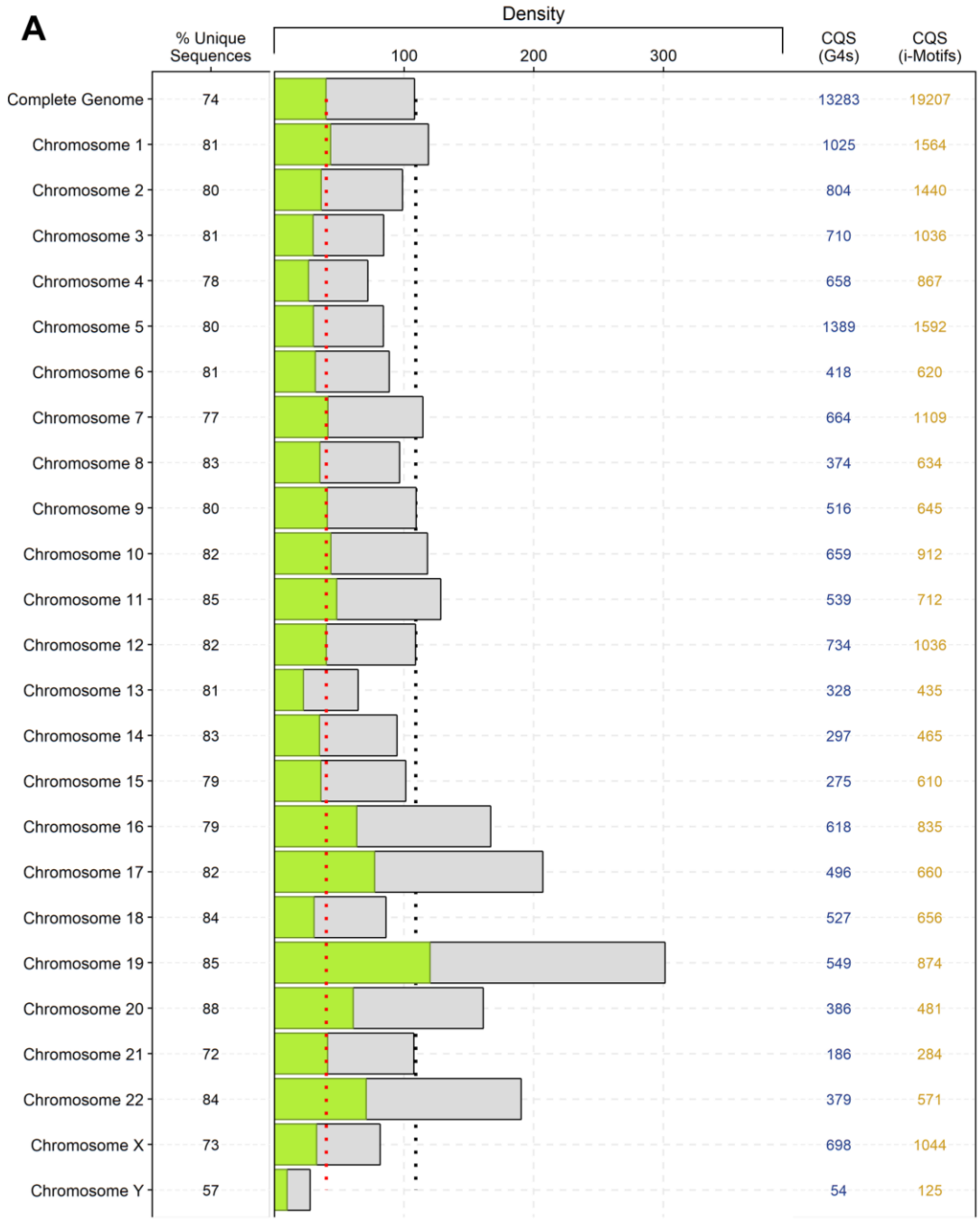
Potential G4 relevance in the genome of humans and other organisms

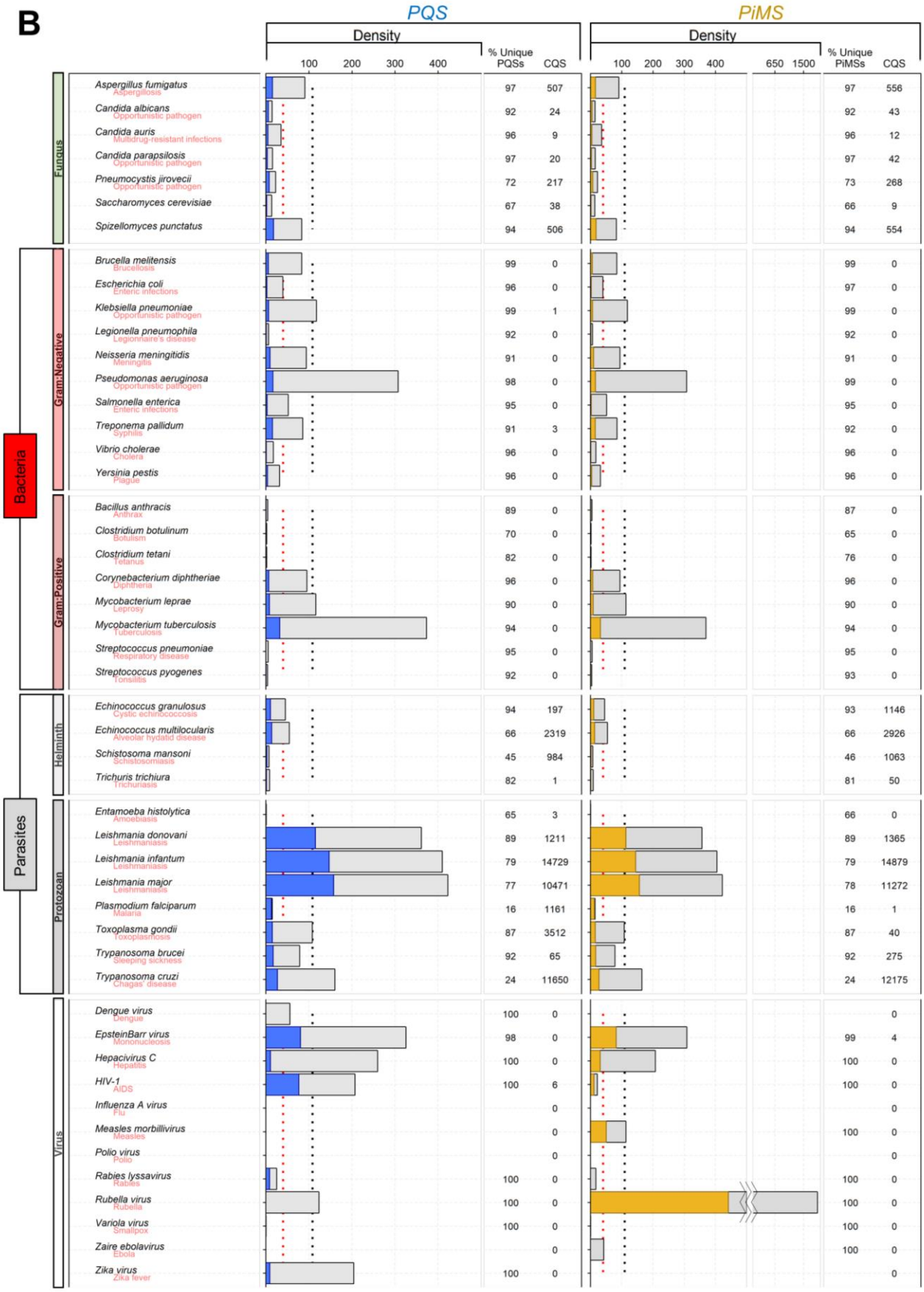
We used G4-iM Grinder's results to analyse macroscopically the human genome. PQS and PiMS densities (per 100000 nucleotides) were analysed, taking into account both the overall and the most probable to form sequences (those that score at least 40; Figure 4, A). In addition, the genomic quadruplex uniqueness and the number of already confirmed G4s and i-Motif structures detected were calculated. The complete average genomic human density was then used as a reference to compare the human chromosomes. The search was then extended to other species that cause mortal and/or morbid pathologies in humans, including: viruses, fungi, bacteria and parasites. In doing so, we wanted to identify the diseases which could potentially be most effectively treated by targeting these therapeutic structures. We decided to use the genomic sequence density as the means to compare chromosomes and genomes because of the huge length differences between them all. Doing so enabled the most practical and simple method to efficiently and clearly compare the prevalence of these potential structures independently of the genomic length.

The combination of applying different scoring criteria and analytical methods to this -first of its class- quadruplex study allowed a wider context of result interpretation.

Depending on the G4-iM Grinder method employed and the scoring criteria used, the human genome potential structure density (both PQS and PiMS) oscillates between 10 and 300 per 100000 nucleotides. Chromosome 19 showed the highest density (with over 3-fold the human average) followed by chromosomes 17, 22 and 16. Chromosome Y, by the contrary, revealed the smallest genomic density and the lowest percentage of unique sequences. In general, sequences found in the human genome present high frequency of repetition (with just 74 % being unique) and high chance of *in vitro* formation, being a third of the total results over 40 in score. In all chromosomes, hundreds to thousands of already confirmed G4 and i-Motifs were detected. These values surpass most other species examined. However, some exceptions exist (Figure 4, B).

Figure 4. Method 2A (M2A) percentage of unique sequences (sequences which have an occurrence frequency of 1 calculated by the formula: [(M2B results which frequency is 1)/(M2A results)]), PQS and PiMS-densities (per 100000 nucleotides, calculated by the formula: [(M2A results)/(Genome length)] x 100000) and number of known-to-form G4 and i-Motif sequences (abbreviated CQS) detected in humans (**A**) and various other organisms (**B**). Grey bars are the genomic PQS density whilst coloured bars are the genomic density filtered by score of at least 40. Black dotted line is the human average density, and red dotted line is the human average density, which scores at least 40.


A
% Unique Sequences
Density
100
200
300
CQS (G4s)
CQS (i-Motifs)
Complete Genome 74 13283 19207
Chromosome 1 81 1025 1564
Chromosome 2 80 804 1440
Chromosome 3 81 710 1036
Chromosome 4 78 658 867
Chromosome 5 80 1389 1592
Chromosome 6 81 418 620
Chromosome 7 77 664 1109
Chromosome 8 83 374 634
Chromosome 9 80 516 645
Chromosome 10 82 659 912
Chromosome 11 85 539 712
Chromosome 12 82 734 1036
Chromosome 13 81 328 435
Chromosome 14 83 297 465
Chromosome 15 79 275 610
Chromosome 16 79 618 835
Chromosome 17 82 496 660
Chromosome 18 84 527 656
Chromosome 19 85 549 874
Chromosome 20 88 386 481
Chromosome 21 72 186 284
Chromosome 22 84 379 571
Chromosome X 73 698 1044
Chromosome Y 57 54 125


B
PQS
PiMS
Density
100 200 300 400
650 1500
% Unique PQSs
CQS
% Unique PiMSs
Bacteria
Parasites
Fungus
Gram-Negative
Gram-Positive
Helminth
Protozoan
Virus
Aspergillus fumigatus Aspergillosis 97 507 97 556
Candida albicans Opportunistic pathogen 92 24 92 43
Candida auris Multidrug-resistant infections 96 9 96 12
Candida parapsilosis Opportunistic pathogen 97 20 97 42
Pneumocystis jirovecii Opportunistic pathogen 72 217 73 268
Saccharomyces cerevisiae 67 38 66 9
Spizellomyces punctatus 94 506 94 554
Brucella melitensis Brucellosis 99 0 99 0
Escherichia coli Enteric infections 96 0 97 0
Klebsiella pneumoniae Opportunistic pathogen 99 1 99 0
Legionella pneumophila Legionnaire's disease 92 0 92 0
Neisseria meningitidis Meningitis 91 0 91 0
Pseudomonas aeruginosa Opportunistic pathogen 98 0 99 0
Salmonella enterica Enteric infections 95 0 95 0
Treponema pallidum Syphilis 91 3 92 0
Vibrio cholerae Cholera 96 0 96 0
Yersinia pestis Plague 96 0 96 0
Bacillus anthracis Anthrax 89 0 87 0
Clostridium botulinum Botulism 70 0 65 0
Clostridium tetani Tetanus 82 0 76 0
Corynebacterium diphtheriae Diphtheria 96 0 96 0
Mycobacterium leprae Leprosy 90 0 90 0
Mycobacterium tuberculosis Tuberculosis 94 0 94 0
Streptococcus pneumoniae Respiratory disease 95 0 95 0
Streptococcus pyogenes Tonsilitis 92 0 93 0
Echinococcus granulosus Cystic echinococcosis 94 197 93 1146
Echinococcus multilocularis Alveolar hydatid disease 66 2319 66 2926
Schistosoma mansoni Schistosomiasis 45 984 46 1063
Trichuris trichiura Trichuriasis 82 1 81 50
Entamoeba histolytica Amoebiasis 65 3 66 0
Leishmania donovani Leishmaniasis 89 1211 89 1365
Leishmania infantum Leishmaniasis 79 14729 79 14879
Leishmania major Leishmaniasis 77 10471 78 11272
Plasmodium falciparum Malaria 16 1161 16 1
Toxoplasma gondii Toxoplasmosis 87 3512 87 40
Trypanosoma brucei Sleeping sickness 92 65 92 275
Trypanosoma cruzi Chagas' disease 24 11650 24 12175
Dengue virus Dengue 100 0 0
EpsteinBarr virus Mononucleosis 98 0 99 4
Hepacivirus C Hepatitis 100 0 100 0
HIV-1 AIDS 100 6 100 0
Influenza A virus Flu 0 0
Measles morbillivirus Measles 0 100 0
Polio virus Polio 0 0
Rabies lyssavirus Rabies 100 0 100 0
Rubella virus Rubella 100 0 100 0
Variola virus Smallpox 100 0 100 0
Zaire ebolavirus Ebola 0 100 0
Zika virus Zika fever 100 0 0

On the one hand, *Leishmania* (and to a less extent the *Trypanosoma* and *Toxoplasma* genus) have very rich quadruplex genomes with many known-to-form G4 and i-Motif sequences within. In *Leishmania major* for example, over 8000 PQS were detected containing the sequence 22Ag (42) with the motif **GGG**TTA. Also, more than 300 PQS containing T30695 and with less frequency T30177 (38), VEGF (42), Scer21 (48), 26gsc (49), Nef8528 (50), IV-1242540 (47), CEB1 (51), A, CC, C and Bc (52), B-raf (53), A3T (49), 96del (41), 27rap (49) and (TG5T)4 (42) were detected. Regarding i-Motifs, cMyb.S (54), cMyc.PY16 (55) , KRC6 (43) and cJun (45) (besides the telomeric i-Motif (44)) were also found. In *T. gondii*, over 3000 sequences containing the known-to-form Ara24-1 (48) with the motif **GGG**TTTA, in addition to C, Bc (52), Chla27 (48) and 93del (41), together with the i-Motifs cMyb.S (54), cJun (45), cMyc.C20T (56) and RAD17.2 (57) were also localized. On the other hand, *Plasmodium falciparum* and *Entamoeba histolytica* (causers of malaria and amoebiasis respectively) displayed very low quadruplex densities because of their high genomic AT content (80.6 and 75.2 % respectively). Still, these sequences identified within *P. falciparum* are the least unique of all analysed as most are different variants of its telomeric sequence, PfTel -with the motif **GGG**TTXA (where X can be any nucleotide). The other helminthic parasites and fungi examined present lower densities that those found in the human genome although all had many known to form G4s and i-Motifs within their genomes.

Gram-positive bacteria display very low genomic quadruplex densities all together. The exceptions are the *Mycobacterium* genus -etiological cause of leprosy and tuberculosis- and the Corynebacterium bacteria, which causes diphtheria. These can surpass and even duplicate the human average. In opposition, Gram-negative bacteria have higher densities for those studied here. *Pseudomonas aeruginosa* is the most outstanding genome in this group with a genome 3-fold denser than its human counterpart. *Brucella melitensis* and *Neisseria meningitides* (causers of brucellosis and meningitis, respectively) follow next in density. Several confirmed known-to-form sequences were also found in *Treponema pallidum*, (40) indicating that G4s are already interesting targets against syphilis.

The viruses analysed display a wide range of unique quadruplex densities. Most of them have different PQS and PiMS densities due to being single stranded genomes. For PQS, The Epstein-Barr virus and HIV present higher densities than the human average whilst Zika, Rubella, Rabies and Hepatitis C viruses have similar or slightly lower densities. Within HIV, the known-to-form sequence PRO1 (58) was located. Other viruses including Ebola, Influenza, Measles and Polio viruses were found to be totally void of PQS. When analysing PiMS, Rubella genome was found to be extremely dense, followed by the Epstein-Barr virus which also presented the known cMyb.S (54) i-Motif. Both viruses have greater densities than the human average whilst Measles, Hepatitis C and HIV viruses have similar or slightly lower densities. In other genomes (including the Polio, Influenza and Zika viruses) no PiMS were found.

DISCUSSION

G4-iM Grinder is a fast, robust and highly adaptable algorithm capable of locating, identifying, qualifying and quantifying quadruplex DNA and RNA structures. These sequences include potential G-quadruplex, i-Motifs and their higher order forms. The adaptation of three scoring systems through machine learning, the structure frequency analysis and the ability to locate already known-to-form G4s and i-Motifs sequences makes G4-iM Grinder's workflow a practical and easy way to find, filter and select the most interesting quadruplex therapeutic targets in a genome. Furthermore, the modular design and the extensive freedom of variable configuration of G4-iM Grinder gives the user full control of what and how these quadruplex are located and analysed.

Using G4-iM Grinder, we examined the human genome to find new high scoring and frequent quadruplex G and C-based structures as potential new targets. We also identified the longest and most probable higher order sequences to form, some of which have already several known-to-form G4 sequences within. The longest of these structures can involve thousands of nucleotides and hundreds of possible PQS combinations. For example, we analysed HoEBR1 (a

recurrent potential higher order quadruplex sequence with good score) and developed the methodology to calculate the best combinations of PQS subunits to form the structure.

A more macroscopic view of the human genome revealed chromosome 19 as the quadruplex densest chromosome and 13, 18 and Y as the least dense ones (with a fall of nearly 66 %). The human genome is still denser and with less unique sequences than most other species examined. However, some parasites and bacteria, such as those in the *Leishmania* and *mycobacterium* genus, present very high densities surpassing by several fold the human average. Other bacteria, like *Pseudomonas aeruginosa*, *Neisseria meningitides* and *Brucella melitensis* are also very rich in potential quadruplex targets, as are the *Trypanosoma* and *Toxoplasma* parasites. In many of these organisms, we identified several sequences that form G4 and i-Motifs *in vitro*. The bacteria and viruses inspected barely presented these known-to-form sequences because these differ from those listed in the sources used to build the known-to-form database. Still the pathological causers of AIDS, hepatitis C, rubella, zika, measles and dengue fever showed very high densities of unique sequences that also exceed the human average. The sum of all these results reflects the great potential G4s have as therapeutic targets against these diseases that currently kill millions worldwide. Other bacteria, parasites and viruses are poorer in or void of quadruplexes and therefore may require less stringent search criteria to find potential targets (for example accepting G or C-runs of length 2).

Future work includes incorporating G4NN and Quadron (when fully developed) as scoring systems. A Shiny application will be developed for G4-iM Grinder and its subsequent result analysis. The quadruplex database will be maintained online to allow external contributions regarding new *in vitro* quadruplex sequences for their identification within G4-iM Grinder results.

DATA AVAILABITITY

The package and all the results can be found through GitHub ("EfresBR/G4iMGrinder"). Instructions on how to install and use the package can be located in the Supplementary information – 8. G4-iM Grinder package. Results and related information can be located in the Supplementary information – 9. Genomes used and results with all methods.

SUPPLEMENTARY INFORMATION

Supplementary information are available online.

FUNDING

This work was supported by The Spanish Ministerio de Economia y Competitividad (CTQ2015- 64275-P).

ACKNOWLEDGEMENTS

The authors thank B. Belmonte, E. Belmonte-Garcia, M. Soto, M. Arévalo, P. Peñalver, S. Heselden, J. L. Mergny and L. Lacroix for their useful insights regarding this topic.

REFERENCES

1. Biffi,G., Tannahill,D., McCafferty,J. and Balasubramanian,S. (2013) Quantitative visualization of DNA G-quadruplex structures in human cells. *Nat. Chem.*, **5**, 182–186.

2. Eddy,J., Vallur,A.C., Varma,S., Liu,H., Reinhold,W.C., Pommier,Y. and Maizels,N. (2011) G4 motifs correlate with promoter-proximal transcriptional pausing in human genes. *Nucleic Acids Res.*, **39**, 4975–4983.

3. Kikin,O., D'Antonio,L. and Bagga,P.S. (2006) QGRS Mapper: a web-based server for predicting G-quadruplexes in nucleotide sequences. *Nucleic Acids Res.*, **34**, W676–W682.

4. Hon,J., Martínek,T., Zendulka,J. and Lexa,M. (2017) pqsfinder: an exhaustive and imperfection-tolerant search tool for potential quadruplex-forming sequences in R. *Bioinformatics*, **33**, 3373–3379.

5. Sahakyan,A.B., Chambers,V.S., Marsico,G., Santner,T., Di Antonio,M. and Balasubramanian,S. (2017) Machine learning model for sequence-driven DNA G-quadruplex formation. *Sci. Rep.*, **7**, 14535.

6. Guédin,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.

7. 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.

8. Arora,A., Nair,D.R. and Maiti,S. (2009) Effect of flanking bases on quadruplex stability and Watson-Crick duplex competition. *FEBS J.*, **276**, 3628–3640.

9. Huppert,J.L. (2005) Prevalence of quadruplexes in the human genome. *Nucleic Acids Res.*, **33**, 2908–2916.

10. Scaria,V., Hariharan,M., Arora,A. and Maiti,S. (2006) Quadfinder: server for identification and analysis of quadruplex-forming motifs in nucleotide sequences. *Nucleic Acids Res.*, **34**, W683–W685.

11. Bagga,P., D'Antonio,L., Kikin,O. and Zappala,Z. QGRS Mapper 2 | G-quadruplex analysis web tool.

12. Dhapola,P. and Chowdhury,S. (2016) QuadBase2: web server for multiplexed guanine quadruplex mining and visualization. *Nucleic Acids Res.*, **44**, W277–W283.

13. Varizhuk,A., Ischenko,D., Tsvetkov,V., Novikov,R., Kulemin,N., Kaluzhny,D., Vlasenok,M., Naumov,V., Smirnov,I. and Pozmogova,G. (2017) The expanding repertoire of G4 DNA structures. *Biochimie*, **135**, 54–62.

14. Agrawal,P., Lin,C., Mathad,R.I., Carver,M. and Yang,D. (2014) The Major G-Quadruplex Formed in the Human BCL-2 Proximal Promoter Adopts a Parallel Structure with a 13-nt Loop in K^+ Solution. *J. Am. Chem. Soc.*, **136**, 1750–1753.

15. 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.

16. Bedrat,A., Lacroix,L. and Mergny,J.-L. (2016) Re-evaluation of G-quadruplex propensity with G4Hunter. *Nucleic Acids Res.*, **44**, 1746–1759.

17. Beaudoin,J.-D., Jodoin,R. and Perreault,J.-P. (2014) New scoring system to identify RNA G-quadruplex folding. *Nucleic Acids Res.*, **42**, 1209–1223.

18. Garant,J.-M., Perreault,J.-P. and Scott,M.S. (2017) Motif independent identification of potential RNA G-quadruplexes by G4RNA screener. *Bioinformatics*, **33**, 3532–3537.

19. Phan,A.T., Kuryavyi,V., Gaw,H.Y. and Patel,D.J. (2005) Small-molecule interaction with a five-guanine-tract G-quadruplex structure from the human MYC promoter. *Nat. Chem. Biol.*, **1**, 167–173.

20. Omaga,C.A., Fleming,A.M. and Burrows,C.J. (2018) The Fifth Domain in the G-Quadruplex-Forming Sequence of the Human *NEIL3* Promoter Locks DNA Folding in Response to Oxidative Damage. *Biochemistry*, **57**, 2958–2970.

21. Marquevielle,J., Kumar,M.V.V., Mergny,J.-L. and Salgado,G.F. (2018) 1H, 13C, and 15N chemical shift assignments of a G-quadruplex forming sequence within the KRAS proto-oncogene promoter region. *Biomol. NMR Assign.*, **12**, 123–127.

22. Arévalo-Ruiz,M., Doria,F., Belmonte-Reche,E., De Rache,A., Campos-Salinas,J., Lucas,R., Falomir,E., Carda,M., Pérez-Victoria,J.M., Mergny,J.-L., *et al.* (2017) Synthesis, Binding Properties, and Differences in Cell Uptake of G-Quadruplex Ligands Based on Carbohydrate Naphthalene Diimide Conjugates. *Chem. - Eur. J.*, **23**, 2157–2164.

23. Guillon,J., Cohen,A., Gueddouda,N.M., Das,R.N., Moreau,S., Ronga,L., Savrimoutou,S., Basmaciyan,L., Monnier,A., Monget,M., *et al.* (2017) Design, synthesis and antimalarial activity of novel bis{ *N* -[(pyrrolo[1,2- *a*]quinoxalin-4-yl)benzyl]-3-aminopropyl}amine derivatives. *J. Enzyme Inhib. Med. Chem.*, **32**, 547–563.

24. Ohnmacht,S.A. and Neidle,S. (2014) Small-molecule quadruplex-targeted drug discovery. *Bioorg. Med. Chem. Lett.*, **24**, 2602–2612.

25. Belmonte-Reche,E., Martínez-García,M., Guédin,A., Zuffo,M., Arévalo-Ruiz,M., Doria,F., Campos-Salinas,J., Maynadier,M., López-Rubio,J.J., Freccero,M., *et al.* (2018) G-Quadruplex Identification in the Genome of Protozoan Parasites Points to Naphthalene Diimide Ligands as New Antiparasitic Agents. *J. Med. Chem.*, **61**, 1231–1240.

26. Sahakyan,A.B., Murat,P., Mayer,C. and Balasubramanian,S. (2017) G-quadruplex structures within the 3′ UTR of LINE-1 elements stimulate retrotransposition. *Nat. Struct. Mol. Biol.*, **24**, 243–247.

27. Petraccone,L., Trent,J.O. and Chaires,J.B. (2008) The tail of the telomere. *J. Am. Chem. Soc.*, **130**, 16530–16532.

28. Petraccone,L., Spink,C., Trent,J.O., Garbett,N.C., Mekmaysy,C.S., Giancola,C. and Chaires,J.B. (2011) Structure and Stability of Higher-Order Human Telomeric Quadruplexes. *J. Am. Chem. Soc.*, **133**, 20951–20961.

29. Vorlícková,M., Chládková,J., Kejnovská,I., Fialová,M. and Kypr,J. (2005) Guanine tetraplex topology of human telomere DNA is governed by the number of (TTAGGG) repeats. *Nucleic Acids Res.*, **33**, 5851–5860.

30. Bauer,L., Tlučková,K., Tóhová,P. and Viglaský,V. (2011) G-quadruplex motifs arranged in tandem occurring in telomeric repeats and the insulin-linked polymorphic region. *Biochemistry*, **50**, 7484–7492.

31. Gehring,K., Leroy,J.-L. and Guéron,M. (1993) A tetrameric DNA structure with protonated cytosine-cytosine base pairs. *Nature*, **363**, 561–565.

32. Kang,C.H., Berger,I., Lockshin,C., Ratliff,R., Moyzis,R. and Rich,A. (1994) Crystal structure of intercalated four-stranded d(C3T) at 1.4 A resolution. *Proc. Natl. Acad. Sci.*, **91**, 11636–11640.

33. Yang,B. and Rodgers,M.T. (2014) Base-Pairing Energies of Proton-Bound Heterodimers of Cytosine and Modified Cytosines: Implications for the Stability of DNA *i* -Motif Conformations. *J. Am. Chem. Soc.*, **136**, 282–290.

34. Leroy,J.-L., Guéron,M., Mergny,J.-L. and Hélène,C. (1994) Intramolecular folding of a fragment of the cytosine-rich strand of telomeric DNA into an *i-motif*. *Nucleic Acids Res.*, **22**, 1600–1606.

35. Geinguenaud,F., Liquier,J., Brevnov,M.G., Petrauskene,O.V., Alexeev,Y.I., Gromova,E.S. and Taillandier,E. (2000) Parallel self-associated structures formed by T,C-rich sequences at acidic pH. *Biochemistry*, **39**, 12650–12658.

36. Bhavsar-Jog,Y.P., Van Dornshuld,E., Brooks,T.A., Tschumper,G.S. and Wadkins,R.M. (2014) Epigenetic Modification, Dehydration, and Molecular Crowding Effects on the Thermodynamics of i-Motif Structure Formation from C-Rich DNA. *Biochemistry*, **53**, 1586–1594.

37. Varizhuk,A., Ischenko,D., Tsvetkov,V., Novikov,R., Kulemin,N., Kaluzhny,D., Vlasenok,M., Naumov,V., Smirnov,I. and Pozmogova,G. (2017) The expanding repertoire of G4 DNA structures. *Biochimie*, **135**, 54–62.

38. Mukundan,V.T., Do,N.Q. and Phan,A.T. (2011) HIV-1 integrase inhibitor T30177 forms a stacked dimeric G-quadruplex structure containing bulges. *Nucleic Acids Res.*, **39**, 8984–8991.

39. Stegle,O., Payet,L., Mergny,J.-L., MacKay,D.J.C. and Huppert,J.L. (2009) Predicting and understanding the stability of G-quadruplexes. *Bioinformatics*, **25**, i374–i1382.

40. Rehm,C. and Hartig,J.S. (2014) In Vivo Screening for Aptazyme-Based Bacterial Riboswitches. In Ogawa,A. (ed), *Artificial Riboswitches*. Humana Press, Totowa, NJ, Vol. 1111, pp. 237–249.

41. Zhang,S., Wu,Y. and Zhang,W. (2014) G-Quadruplex Structures and Their Interaction Diversity with Ligands. *ChemMedChem*, **9**, 899–911.

42. Tran,P.L.T., Largy,E., Hamon,F., Teulade-Fichou,M.-P. and Mergny,J.-L. (2011) Fluorescence intercalator displacement assay for screening G4 ligands towards a variety of G-quadruplex structures. *Biochimie*, **93**, 1288–1296.

43. Manzini,G., Yathindra,N. and Xodo,L.E. (1994) Evidence for intramolecularly folded i-DNA structures in biologically relevant CCC-repeat sequences. *Nucleic Acids Res.*, **22**, 4634–4640.

44. Lannes,L., Halder,S., Krishnan,Y. and Schwalbe,H. (2015) Tuning the pH Response of i-Motif DNA Oligonucleotides. *ChemBioChem*, **16**, 1647–1656.

45. Saxena,S., Bansal,A. and Kukreti,S. (2008) Structural polymorphism exhibited by a homopurine·homopyrimidine sequence found at the right end of human c-jun protooncogene. *Arch. Biochem. Biophys.*, **471**, 95–108.

46. Gurung,S.P., Schwarz,C., Hall,J.P., Cardin,C.J. and Brazier,J.A. (2015) The importance of loop length on the stability of i-motif structures. *Chem. Commun.*, **51**, 5630–5632.

47. 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 Comput. Biol.*, **6**, e1000861.

48. Tran,P.L.T., Mergny,J.-L. and Alberti,P. (2011) Stability of telomeric G-quadruplexes. *Nucleic Acids Res.*, **39**, 3282–3294.

49. Stegle,O., Payet,L., Mergny,J.-L., MacKay,D.J.C. and Huppert,J.L. (2009) Predicting and understanding the stability of G-quadruplexes. *Bioinformatics*, **25**, i374–i1382.

50. Perrone,R., Nadai,M., Poe,J.A., Frasson,I., Palumbo,M., Palù,G., Smithgall,T.E. and Richter,S.N. (2013) Formation of a Unique Cluster of G-Quadruplex Structures in the HIV-1 nef Coding Region: Implications for Antiviral Activity. *PLoS ONE*, **8**, e73121.

51. Adrian,M., Ang,D.J., Lech,C.J., Heddi,B., Nicolas,A. and Phan,A.T. (2014) Structure and Conformational Dynamics of a Stacked Dimeric G-Quadruplex Formed by the Human CEB1 Minisatellite. *J. Am. Chem. Soc.*, **136**, 6297–6305.

52. 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**, 677–677.

53. Wei,D., Todd,A.K., Zloh,M., Gunaratnam,M., Parkinson,G.N. and Neidle,S. (2013) Crystal Structure of a Promoter Sequence in the *B-raf* Gene Reveals an Intertwined Dimer Quadruplex. *J. Am. Chem. Soc.*, **135**, 19319–19329.

54. Brazier,J.A., Shah,A. and Brown,G.D. (2012) I-Motif formation in gene promoters: unusually stable formation in sequences complementary to known G-quadruplexes. *Chem. Commun.*, **48**, 10739–10741.

55. Simonsson,T., Pribylova,M. and Vorlickova,M. (2000) A Nuclease Hypersensitive Element in the Human c-myc Promoter Adopts Several Distinct i-Tetraplex Structures. *Biochem. Biophys. Res. Commun.*, **278**, 158–166.

56. Dai,J., Hatzakis,E., Hurley,L.H. and Yang,D. (2010) I-Motif Structures Formed in the Human c-MYC Promoter Are Highly Dynamic–Insights into Sequence Redundancy and I-Motif Stability. *PLoS ONE*, **5**, e11647.

57. Rogers,R.A., Fleming,A.M. and Burrows,C.J. (2018) Rapid Screen of Potential i-Motif Forming Sequences in DNA Repair Gene Promoters. *ACS Omega*, **3**, 9630–9635.

58. 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.

Supplementary Information for

G4-iM Grinder: When size and frequency matter. G-Quadruplex, i-Motif and higher order structure search and analysis tool.

Efres Belmonte-Reche* and Juan Carlos Morales.

Contents table:

1. G4-iM Grinder's search and analyser algorithm S2
2. Variables, predefined values and examples S4
3. G-Quadruplex and G4-iM Grinder S6
4. i-Motifs and G4-iM Grinder S9
5. Scoring models and their adaptations S10
    a. cGcC
    b. PQSfinder
    c. G4hunter
    d. Final Score evaluation
    e. Scoring system for i-Motifs
6. Explanation of other Variables S18
    Complementary and *DNA*
    Composition Variables
        LoopSeq
        KnownQuadruplex
        KnownNOTQuadruplex
7. G4-iM Grinder performance S20
8. G4-iM Grinder package S22
9. Genomes used and results with all methods S25
10. References S27

1. G4-iM Grinder's search and analyser algorithm

The algorithm code is written in R and divided into several parts. After a setup is executed, the genomic sequence is analysed and converted into an acceptable format. Then, the complementary strand (if *Complementary* = TRUE) is created. The core of the algorithm is divided in 4 parts, the first two being the search engines (Method 1A and 1B, M1A and M1B respectively), and the other two analysers (Method 2 for overlapping and size-dependent search, and Method 3 for non-overlapping size-independent examination, M2A and M3A respectively).

M1A reads and locates all the possible nucleotide runs in the genomic sequence. It starts by identifying non-overlapping perfect runs (with no bulges) and then proceeds by finding imperfect ones in a greedy manner, meaning that it will look for and preferably accept bigger runs than smaller ones. To work, it requires four variables (examples can be found in the supplementary information – S4.Variables, predefined values and examples) which are: *RunComposition*, *BulgeSize*, *MaxRunSize* and *MinRunSize.* M1A output is then passed to M1B, which identifies the relationship between the runs. The method starts by trying to relate runs with their direct neighbours and if unsuccessful, it will expand the search further until *MaxLoopSize* is reached. If this fails, M1B will calculate if reducing the flanking run sizes whilst within the *MinRunSize* acceptable range can solve the problem. This process is dependent on: *MaxLoopSize* and *MinLoopSize*.

M1B data is then analysed by Method 2 (M2A) if the option is selected. This method will join several linked runs to yield the final structures if they comply with the user-defined parameters: *MaxNRuns*, *MinNRuns*, *MaxPQSSize*, *MinPQSSize* and *MaxIL* (maximum total number of Bulges accepted per sequence). This is repeated for each linked run so overlapping PQS are all identified together with their position. If desired, *MaxNRuns* can be set to zero to cancel its use in the search and hence accept sequences with more than four runs.

If potential higher-order structures are to be detected, Method 3 (M3A) will select from the M1B data all the runs that have no link with the previous yet are linked with the proceeding

ones. These runs will be considered leaders, to which the algorithm will build structures by following the consecutive links. Contrary to M2A, M3A will continue over the *MaxPQSSize* and *MaxNRuns* limits. Hence, it will continue constructing a potential higher order sequence until a non-linked run is added or the end of the sequence is reached. This method depends on the *MinNRuns* and *MinPQSSize* variables.

Once all M2A and M3A results have been identified, G4-iM Grinder will count each structure with the same sequence to calculate its frequency in the genome. These results are stored in a new table called M2B and M3B respectively.

If the computer can and the user allows (by ceding workers through the *NCores* variable), several functions have been given the capacity of applying parallelized computation. This results in a faster analysis.

2. Variables, predefined values and examples.

Table 1: Variables and predefined values for G4-iM Grinder.

Category	Variable Name	Predefined value	Explanation
Function	*NCores*	1	Nº of Cores to cede to the function for parallel computation.
	Verborrea	TRUE	Allow the function to update the user with its progress.
Sequence	*Complementary*	TRUE	Analyse the complementary strand of the imputed sequence.
	DNA	TRUE	Identifies whether the Sequence is a DNA sequence. If false, it will assume its RNA. Useful for complementary conversion and *KnownQuadruplex* and *KnownNOTQuadruplex*.
For M1A	*RunComposition*	"G"	Any character can be selected to determine which nucleotide forms the runs. G for G4, C for i-Motifs, tolerates T, U and A or anything else.
	BulgeSize	1	Number of tolerated non-*RunComposition* nucleotides within the runs, as to allow runs with bulges.
	MaxRunSize and *MinRunSize*	5 -3	Determines the maximum and minimum numbers of the *RunComposition* nucleotides to be considered a run.
For M1B	*MaxLoopSize* and *MinLoopSize*	10 - 0	Maximum and Minimum number of nucleotides required to assume relationship between runs.
For M2A	*Method2*	TRUE	To apply method 2. Search for structures with defined size and runs. Will also give frequency counts of each structure detected (M2B).
	MaxNRuns and *MinNRuns*	0 - 4	Maximum and Minimum number of linked runs which are necessary to form a structure. *MaxNRuns* can be set to 0 to avoid using it as a limiting condition for structure formation (to rely just on structure size).
	MaxPQSSize and *MinPQSSize*	33 - 15	Maximum and Minimum number of nucleotides which can be part of the final structures.
	MaxIL	3	Total maximum numbers of bulges allowed for the whole structure.
For M3A	*Method3*	TRUE	To apply method 3. Search for structures with unrestricted size and numbers of runs. Useful for searching higher forming structures. Depends on variables: *MinNRuns* and *MinPQSSize*.
Quantification	*LoopSeq*	G, C, T, A	Quantification of each element of *LoopSeq* in all the PQS found (as a % of total length of the structure). Multicharacter values are accepted (like GGG or TTA)
Qualification	*KnownQuadruplex*	TRUE	Identification of known-to-form-quadruplex (DNA or RNA, G4s or i-Motifs).
	KnownNOTQuadruplex	FALSE	Identification of known-NOT-to-form-quadruplex (DNA or RNA, G4s or i-Motifs).
Scoring	*G4Hunter*	TRUE	To apply G4-hunter evaluation system to the detected final sequences.(1) The results are transformed into a percentage, from 100 to -100 %.
	PQSfinder	TRUE	To apply an adaptation of the PQSfinder Scoring system described by Hon et al.(2) It is dependent on *RunComposition* = G or C.
	cGcC	FALSE	To apply the cGcC system adopted from Beaudoin *et al*.(3) for RNA sequences. It is dependent on *RunComposition* = G or C.
	WeightParameters	50, 50, 0	Weighted values of each scoring system as to calculate the final score through weighted averages. Order of vector: G4hunter, PQSfinder & cGcC.
	FreqWeight	0	Weight of the structure frequency in the final score value for scoring calculations of M2B and M3B.

An additional nine parameters can be modified for PQSfinder analysis.

Variable Examples:

M1A examples:

Variables	Sequence	Result
RunComposition = "G" *MinRunSize* = 3 *BulgeSize* = 0	GGG GG GGAG	Accepted Non-accepted Non-accepted
RunComposition: "G" *MinRunSize*: 4 *BulgeSize*: 1	GGGG GGAG GGAGG	Accepted Non-accepted Accepted
RunComposition: "C" *MinRunSize*: 3 *BulgeSize*: 2	CCC CTTCC CTC	Accepted Accepted Non-accepted

M1B examples:

Variables	Sequence	Result
MaxLoopSize: 10 *MinLoopSize*: 1	GGGTGGG GGGTTTTATTTTACGGG	Linked Unlinked
MaxLoopSize: 7 *MinLoopSize*: 3	GGGTGGG GGGTTTTAGGG	Unlinked Linked
MaxLoopSize: 4 *MinLoopSize*: 0	GGGGGG GGGTTTTAGGG	Linked Unlinked

M2A examples:

Variables	Sequence	Result
MinNRuns: 4 *MaxPQSSize*: 20 *MinPQSSize*: 15	GGGTGGGTGGGTTTTAGGG GGGTTTTAGGGTTTTAGGG GGGTTTTAGGGTTTTAGGGTTTTAGGG	PQS Not-PQS Not-PQS
MinNRuns: 2 *MaxPQSSize*: 10 *MinPQSSize*: 8	GGGTGGG GGGTTTAGGG GGGTGGGGGG	Not-PQS PQS PQS

3. G-Quadruplex and G4-iM Grinder

392 sequences from the supplementary material of G4Hunter were evaluated to calculate the classification performance of G4-iM Grinder's Method 2 using predefined parameters. This list of sequences is composed of 94 genomes which do not form G4 and 298 that do.

When the algorithm was applied to the non-forming G4 sequences, 93 of 94 sequences were not recognised as PQS. Only X3CGC with the sequence **GGGCGCGGGCGCGGGCGCGGG** was falsely recognised as a potential quadruplex with fairly high scores given by G4Hunter and PQSfinder (mean score = 48).

When applied to the G4-forming sequences, 233 out of the 298 were correctly recognised with G4-iM Grinder's predefined values whilst 65 were not. The 5 main reasons why these were not detected are as follows (Table 2):

1) *MinNRuns* < 4: Predefined search values require the presence of at least four G-runs to accept a sequence as a PQS. Therefore, G4s that are intermolecular (with less than 4 G-runs) are not recognised when analysed independently. However, most of these sequences have been found as part of bigger structures.

2) *MinRunSize* < 3: Predefined search values require 3 guanine G-runs to be detected. Hence, neither G-runs of size 2 nor the structures they form were recognised.

3) *BulgeSize* > 1 & *MaxIL* > 3: Predefined search values require sequences to have just 1 bulge per G-run and no more than 3 total bulges per sequence.

4) *MaxLoopSize* > 10: Predefined search values require sequences to have loops of 10 nucleotides or smaller.

5) *MaxPQSSize* > 33: Predefined search values require sequences to have a total length of no more than 33 nucleotides. However this reason can be avoided by analysing the sequences also with the size unrestricted Method 3 (M3).

These search variables can be modified to include these characteristics in any G4-iM Grinder analysis.

This analysis with the predefined variables and its confusion matrix is:

		Real Condition			
		Positive	Negative		
Predicted Condition	Positive	233	1	234	Prevalence = 76.02%
	Negative	65	93	158	Accuracy = 83.16%
		298	94	**392**	

Sensitivity (TPR) =	78.19%	Precision (PPV) =	99.57%
False negative rate (FNR) =	21.81%	False discovery rate (FDR) =	0.43%
Fall-out (FPR) =	1.06%	False omission rate (FOR) =	41.14%
Specificity (SPC) =	98.94%	Negative predicted value (NPV) =	58.86%

The scores (mean of G4Hunter and G4-iM Grinder's PQSfinder) of the 233 Sequences were then examined. The mean score of all these structures was of 52.96. Of these structures, 75 % presented a score of 48 or more, and 98 % (228 of 233) scored more than 40. Hence, the filter used in this article to quantify the most probable to form PQS (Score ≥ 40) scored at least the same as 97.9 % of all verified to form sequences examined.

Table 2: Verified G4 sequences not recognised by G4-iM Grinder using predefined search parameters. The reason why the Sequence was not recognised is given in the third column.

Name	Sequence	Reason
C	GGAGGGTGGATGG	Min G-run size < 3
Cc	AGAGGGTAGATGG	Nº G-runs < 4
B	GGGGGATGCGGGGG	Nº G-runs < 4
Bc	AGGAGATGCAGGAG	Nº G-runs < 4
TBA	GGTTGGTGTGGTTGG	Min G-run size < 3
15G1	GGTTGGTTAGGTTGG	Min G-run size < 3
15GT	GGTTGGTTTGGTTGG	Min G-run size < 3
Nef8528	GAGGAGGAGGTGGGT	Nº G-runs < 4
93del	GGGGTGGGAGGAGGGT	Min G-run size < 3 / Nº G-runs < 4
16G1	GGTTGGTTTTGGTTGG	Min G-run size < 3
Nef8624	GGGGGGACTGGAAGGG	Min G-run size < 3 / Nº G-runs < 4
A	GGATGGGGTGGGGAGG	Min G-run size < 3
Ac	AGATGGAGTGGAGAGG	Min G-run size < 3
Bom17	GGTTAGGTTAGGTTAGG	Min G-run size < 3
CEB1	AGGGGGGAGGGAGGGTGG	Min G-run size < 3
18gtel2	AGGTTAGGTTAGGTTAGG	Min G-run size < 3
PS2,M	GTGGGTAGGGCGGGTTGG	Min G-run size < 3
PRO1	TGGCCTGGGCGGGACTGGG	Min G-run size < 3
UTX	GCCGGGCGGGGAGGGGGGGTCA	Nº G-runs < 4
c-kit*	GGCGAGGAGGGGCGTGGCCGGC	Min G-run size < 3
HPV42	GGGACTATGGGTAACGGGGGGG	Nº G-runs < 4
12668310-PC12-16	AGAGTGGGGGGGATGTAGGTGGGTT	Min G-run size < 3 / Nº G-runs < 4
12668310-PC12-9	AGTGGGGGTAGGGGATAGGGTAGGC	Min G-run size < 3
12668310-PC12-10	GCTGGGGTGTTGGGTGTGGGGGTGA	Nº G-runs < 4
25DDX	GGGCGGGAUAGAGAGCGUGGGCGGG	Loop > 10
12668310-PC12-23	GGGTGTGAGAGGTTGAGGGGGTTCG	Min G-run size < 3 / Nº G-runs < 4
12668310-PC12-17	GGTTGGATGTAAGGTTGGAGGGGGG	Min G-run size < 3
12668310-PC12-4	TATGGGGGTGGGTCAGGTTTCGGTA	Min G-run size < 3
12668310-PC12-2	TGAGGGTCTAGGGTGGTGGGGTGGA	Min G-run size < 3
12668310-PC12-3	TGATGGATGTGGGGATGCGGGGGCG	Min G-run size < 3 / bulges > 1
12668310-PC12-15	TGGGTAGGTTCGAGGGGTGGGTGTG	Min G-run size < 3
12668310-PC12-1	TGGTTGGGGATAGAGGTGGGTGTTT	Min G-run size < 3
H-Bi-G4	GGGACGTAGTGGGGGGACGTAGTGGG	Nº G-runs < 4
AGRO100	GGTGGTGGTGGTTGTGGTGGTGGTGG	Total Bulges > 3
VNTR6-1	GGGGTAGGTGGGGATCTGTGGGATTGG	Min G-run size < 3
f1E1t	GGGTGGGTTTTTTTTTTTTTTTTGGGTGGG	Loop > 10
Nef8547	GGTCTTAAAGGTACCTGAGGTCTGACTGG	Min G-run size < 3
21704505-FADGDH-1	TCCGGGGGGCTGGGCAGGGGGGTAACTTTC	Nº G-runs < 4
CSBIIWT	GAAGCGGGGGAGGGGGGGUUUGGUGGAAAU	Min G-run size < 3
32T1H1	GGGTGGGTTTTTTTTTTTTTTTTTTTGGGTGGG	Loop > 10
Cppt2	TTTTAAAAGAAGGGGAGGAATAGGGGATATGA	Nº G-runs < 4
f3E3t	GGGTTTGGGTTTTTTTTTTTTTTTTGGGTTTGGG	Loop > 10
15025912-NS5B-18	GGGGTAGGATAGGGTNTGGAAGGAGGTGCCCCGT	Min G-run size < 3
7542922-HIV-1-RT-3	CAGGCGTTAGGGAAGGGCGTCGAAAGCAGGGTGGG	Loop > 10
f1K1t	GGGTGGGTTTTTTTTTTTTTTTTTTTTGGGTGGG	Loop > 10 / Max PQS size > 33
cellobiose-3	GTCAAGGTGGGTGGGTGGGGTTGGTTGTTGTTTTGA	Min G-run size < 3
LysG1	TGGGACCATTGAGGGTGGGAAATTGGACAATGGGGA	Loop > 10 / Max PQS size > 33
LysG3	CGGGGTCCGAGGGGATTCCTAAGGGGGTTCTGGGGA	Max PQS size > 33
16518777A-Ricin-4	GGGGGAGGACGCGTAGTGGGGGGCCCATGGTTGTGTGG	bulges > 1 / Max PQS size > 33
38T1N1	GGGTGGGTTTTTTTTTTTTTTTTTTTTTTTTGGGTGGG	Loop > 10 / Max PQS size > 33
20971648-rHuEPO-a-Ma-2	GATTGAAAGGTCTGTTTTTGGGGTTGGTTTGGGTCAATA	Min G-run size < 3
f3K3t	GGGTTTGGGTTTTTTTTTTTTTTTTTTTTTTTGGGTTTGGG	Loop > 10 / Max PQS size > 33
21531729-CD16acMet-7	ATCACGTGGTGGGCAAATAACCGGTTGGGGTGGGTCGAGG	Min G-run size < 3
21531729-CD16acMet-3	GAGCGGGGACGAACACATATGGGGAAGTGGCTTGGGGTGG	Min G-run size < 3 / Loop > 10 / Max PQS size > 33
21531729-CD16acMet-2	GAGTGCGTAATGGTACGATTTGGGAAGTGGCTTGGGGTGG	Min G-run size < 3 / Loop > 10
16518777B-Ricin-3	GGAGGCGCGATGTAGGTATGTGAGGGCGGCGCGGTGGGCG	Loop > 10 / Max PQS size > 33 / Max Bulges > 3
15984861-NeuropeptideY-12	TGTGAAGGGGGGTACATGACGGGGACTGGCCGGACTACAG	Min G-run size < 3
Cppt1	TTTTAAAAGAAAAGGGGGGATTGGGGGGTACAGTGCAGGGG	Min G-run size < 3
SMG4T6	TCACAGGGGTTTTTTGGGGTTTTTTGGGGTTTTTTGGGGACAA	Max PQS size > 33
f1S1t	GGGTGGGTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTGGGTGGG	Loop > 10
X-106443	GGGTCCTCCAAGGGGTAAAACTTACATGGGATGGTGGGGTCACATGGG	Loop > 10 / Max PQS size > 33
f3S3t	GGGTTTGGGTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTGGGTTTGGG	Loop > 10
14744035-HIV-1NucleocapsidProtein-4	TCGAGGGGTGTGCAAGGCGGGTCAACGGGCCTTATTTGGTGCTTAGGTA	Min G-run size < 3 / Max PQS size > 33
GAG	AGCGGGGGAGAAUUAGAUAAAUGGGAAAAAAUUCGGUUAAGGCCAGGGGGAAA	Min G-run size < 3 / Loop > 10 / Max PQS size > 33
CLN003	GGAGGGAAAAGTTATCAGGCTGGATGGTAGCTCGGTCGGGGTGGGTGGGTTGGCAAGTCT	Loop > 10 / Max PQS size > 33

4. i-Motif and G4-iM Grinder

100 confirmed i-Motifs were documented from literature and referenced. pH 5, 6, 6.5 and 7 melting temperatures (Tm) were also recorded if stated, although experimental conditions were disregarded. These 100 i-Motifs were analysed by G4-iM Grinder. 77 of the detected sequences were found using G4-iM Grinder's predefined parameters. 4 were not detected because they present 3 or fewer C-runs (*MinNRuns* < 4) and hence are intermolecular i-Motifs, or are too small (*MinPQSSize* < 15, Table 3). 9 more were detected with the size independent method 3 (M3) as they surpass the max. number of nucleotides (*MaxPQSSize* > 33) for M2. 11 i-Motifs have C-runs of size 2 (*MinRunSize* = 2), and 7 longer loops (*MaxLoopSize* > 10) than accepted.

G4-iM Grinder's i-Motif Sensitivity and False Negative Rate with predefined values is 77 and 23 %. However, all these search variables can be modified to locate all these characteristics in any G4-iM Grinder analysis.

Table 3: Verified i-Motifs sequences not recognised by G4-iM Grinder using predefined search parameters. The reason why the Sequence was not recognised is stated in the comments column.

Name	PQS	Length	G4Hunter	pqsfinder	Comments
nMyc	CCCCCTGCATCTGCATGCCCCCTCCCACCCCC	32	-54	-74	Loop > 10
nMyc.mod	CCCCCTGCATCTTTTTGCCCCCTCCCACCCCC	32	-54	-74	Loop > 10
SMC01	CCCTTGCTATCCCTGTCCTGCCTCGCCCTTGGTCATGAACCCC	43	-26	-49	Loop > 10
SMC01.4X4C	CCCCTTGCTATCCCCTGTCCTGCCTCGCCCCTTGGTCATGAACCCC	46	-36	-56	Loop > 10
SMC01.4X4C.AA	CCCCTTGCTATCCCCTGTAATGAATCGCCCCTTGGTCATGAACCCC	46	-32	-51	Loop > 10
SMC04	CCTGCCTCGCCCTTGGTCATGAACCCC	27	-26	-48	Loop > 10
bcl2.1	CCCCGCTCCCGCCCCCTTCCTCCCGCGCCCGCCCC	35	-57	-69	MaxPQSSize > 33
C3T888	CCCTTTTTTTTCCCTTTTTTTTCCCTTTTTTTTCCC	36	-25	-43	MaxPQSSize > 33
FEN1	CCCCACTCCACCCACACCAGGTCCCCGCAGGCCCCTGCTCCC	42	-40	-44	MaxPQSSize > 33
HIF1a.L	CCCGCCCCCTCTCCCCTCCCCGCGCGCCCGAGCGCGCCTCCGCCCTTGCCCGCCCCC	57	-48	-60	MaxPQSSize > 33
hTERT.2	CCCCGCCCCGTCCCGACCCCTCCCGGGTCCCCGGCCCAGCCCCC	44	-53	-72	MaxPQSSize > 33
MGMT	CCCCAGCTCCGCCCCCGCGCGCCCCGGCCCCGCCCCC	37	-57	-65	MaxPQSSize > 33
PDGF.A	CCGCGCCCCTCCCCCGCCCCCGCCCCCGCCCCCCCCCCCCC	41	-78	-75	MaxPQSSize > 33
RAD21	CCCCACCCCCTCCCCCGACCCTTTTCCCCTCCCC	34	-71	-78	MaxPQSSize > 33
RAD54L	CCCCGCCCCTCCCCCGCCACCCCCGCCCCCGCCCCGCCCC	40	-77	-82	MaxPQSSize > 33
c-Myb	CCTCCTCCTCCTTCTCCTCCTCCTCCGTGTCCTCCTCCTCC	41	-29	-48	Min C-run size < 3
cMyc.2	CCTACCCTCCCTACCC	16	-48	-58	Min C-run size < 3
cMyc.2.mod	CCTACCCTTCCTACCC	16	-41	-54	Min C-run size < 3
cMyc.PY27.1	CCTTCCCCACCCTCCCCACCCTCCCC	26	-67	-68	Min C-run size < 3
HCSIII	CCATTCCATTCCTTTCC	17	-24	-44	Min C-run size < 3
hTel.mod	CCTTTCCTTTACCTTTCC	18	-22	-43	Min C-run size < 3
hTERT.1	CCGGGCCCTCCCAGCCCCTCCCC	23	-48	-65	Min C-run size < 3
RET	CCGCCCCCGCCCCGCCCCGCCCC	23	-74	-75	Min C-run size < 3
SMC02	CCCTTGCTATCCCTGTCCTGCC	22	-27	-48	Min C-run size < 3
SMC03	CCCTGTCCTGCCTCGCCC	18	-33	-52	Min C-run size < 3
XRCC2	CCCACCGGCGGCCTTGTTCCCATCTCCCTCACTCCCAACCC	41	-29	-45	Min C-run size < 3
cMyb.S	CCTCCTCCTCCTCC	14			MinPQSSize < 15
Rb	CCGCCCAAAACCCCCC	16			Nº C-runs < 4
cJun	CCCCCTCCCCCTCCCCC	17	Unused for analysis		Nº C-runs < 4
UDG.2	CCCCTAAGGGGCAGGAACTTTTCTTCCCAGCCCCCTCCCCCGACCCC	47			Loop > 10
SG20c	CCCCCAATTAATTCCCCC	18			Nº C-runs < 4

5. Scoring models and their adaptations

G4-iM Grinder is capable of evaluating potential quadruplex sequences using several published and contrasted scoring algorithms. cGcC was adapted and PQSfinder was upgraded to fit G4-iM Grinder's workflow. G4hunter was directly implemented from the original article's supplementary material.

a) cGcC

cGcC was designed by Beaudoin *et al*. (3) to evaluate RNA PQS by using the formula:

$$cGcC = \frac{\sum_{1:25}^{n}(Gs(i) \times 10 \times i) + 1}{\sum_{1:25}^{n}(Cs(i) \times 10 \times i) + 1}$$

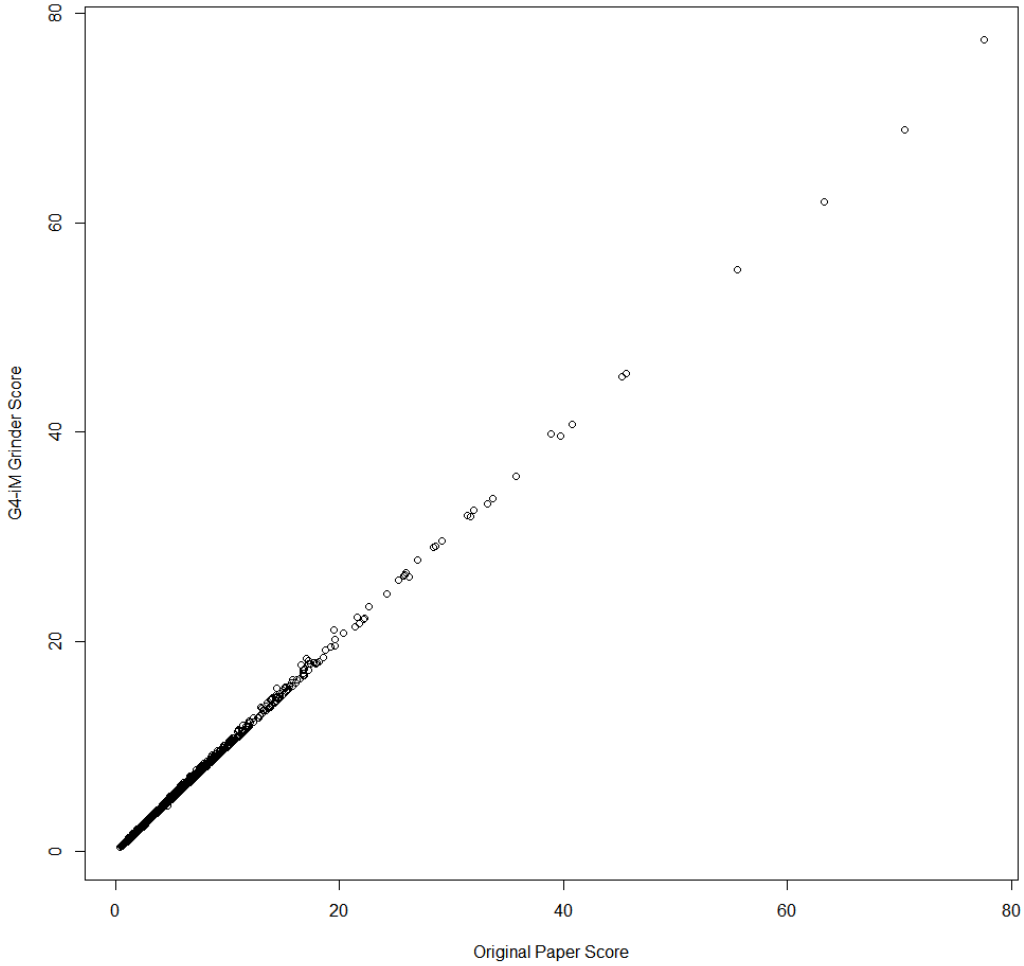
Where Gs(i) or Cs(i) are the number of G or C elements which compose a run, and i is the number of runs. A range of 50 nucleotides are analysed before and after the potential quadruplex sequence and the presence of complementary base nucleotides diminishes the final score. An example is GGGCCCGGG which would score:

$$G4RNA\ (G) = \frac{(\mathrm{GGG})\ 3\ \times\ 10\ \times\ 2\ +\ (\mathrm{GG})\ 2\ \times\ 10\ \times\ 4\ +\ (\mathrm{G})\ 1\ \times\ 10\ \times\ 6}{(\mathrm{CCC})\ 3\ \times\ 10\ \times\ 1\ +\ (\mathrm{CC})\ 2\ \times\ 10\ \times\ 2\ +\ (\mathrm{C})\ 1\ \times\ 10\ \times\ 3} = 2$$

Given the nature of this scoring system, cGcC results are uncontained in the -100 to 100 range and therefore we recommend using cGcC independently of other score systems in the final global scoring evaluation (modulated by the *WeightedParameters* variable).

The >2000 sequences and results found in G4RNA's supplementary material (4) were compared to G4-iM Grinder's cGcC function. The results show an almost perfect correlation between both scores (Figure 1).

Figure 1: comparison of cGcC results between the G4RNA article and those obtained by G4-iM Grinder cGcC function.



b) PQSfinder:

The original PQSfinder scoring system (2) is based on a modulated formula which analyses quadruplex characteristics such as runs, bulges and loops.

Several modifications were made to adapt it to G4-iM Grinder and to allow greater freedom of analysis. This is because the original function does not allow i-Motif, higher order sequence, or evaluation of sequences with more than four G runs. Low scoring sequences are also not returned.

To adapt the original formula, the appendix of the G run bulges penalisation variables *Fb* (bulge length penalisation factor), *Lbi* (the length of the i-th bulge) and *Eb* (bulge length exponent)

were simplified to 1. The number of tetrads (Nt) variable was substituted by the average run size (mNt, excluding from this calculation the bulges) to better analyse longer sequences. Also, the bulge penalisation factor was modulated by multiplying it by the relationship between the minimal and the actual Nº of Runs. This was done to avoid scoring distortions in larger sequences due to the characteristic dependency on the absolute number of bulges. Additionally, all segments of the formula were modulated by supplement and exponential constants. The formula used was:

$$PQSfinder = \text{Tetrad Value} - \text{Bulge Value} - \text{Loop Value}$$

$$\text{Tetrad Value} = \left((mNt - 1)Bt + Ts\right)^{Et}$$

$$Bulge\ Value = \left(\left(Nb \times (Pb + 1) \times \frac{MinNoR}{NoR}\right) + Is\right)^{Ei}$$

$$\text{Loop Value} = \left(\left(\frac{Length - (NoR \times mNt) - Nb}{NoR - 1} \times Fm\right) + Ls\right)^{Em}$$

Where *mNt* is the mean run size, *Nb* is the number of bulges within the runs, *Length* is the number of nucleotides in the potential quadruplex and *NoR* the number of runs in the sequence.

G4-iM Grinder's PQSfinder formula required a re-evaluation to find values that give the nearest score to that of the original PQSfinder, and hence to their *in-silico* G4 formation criteria. The formula was assessed in the non-chromosomal human DNA sequence using both the original PQSfinder and G4-iM Grinder's PQSfinder whilst accepting only structures with 4 runs (PQSfinder only locates these). 11359 PQS were used to train the parameters (as another 7053 had to be discarded due to the scoring restrictions of the original PQSfinder) by developing and using a machine learning algorithm capable of finding the best values which give the minimal score differences. The range of examination and the best values found are stated in Table 4.

Table 4: PQSfinder variables and their values for both the original and G4-iM Grinder's PQSfinder algorithms. The tested optimization range is displayed in the last column and results of the process are shown underneath.

	Name	Description	Original Values	Recommended New Values	Range examined
Original Constants	*Bt*	Tetrad stacking bonus constant	40	14	10 to 50
	Pb	Bulge penalisation constant	20	17	10 to 30
	Fm	Loop length penalisation constant	6.6	3	0 to 10
	Em	Loop length exponential constant	0.8	1	0 to 2
New Constants	*Ts*	Tetrad supplement constant		4	0 to 20
	Is	Bulge supplement constant		-19	-20 to 20
	Ls	Loop supplement constant		-16	-20 to 20
	Et	Tetrad exponential constant		1	0 to 2
	Ei	Bulge exponential constant		1	0 to 2
	ET	Total formula exponential constant		1	0 to 2

Results	Mean Difference	4.92
	R^2	0.89
	Percentiles of PQS population	50 % - < 3 Scoring difference 75 % - < 6 Scoring difference 95 % - < 16 Scoring difference

The mean score difference for the same PQS between algorithms was of 4.92 (Original score = G4-iM Grinder score ± 4.92). 50 % of the 11359 PQS analysed fell inside a ± 3 window, 75 % of ± 6 and 95 % of ± 16 (Figure 2). Punctual PQS scoring variations occurred due to different G-run, Loop and bulge definitions, as well as formula variances. However, similarities with the original scoring system are sufficient (R^2 = 0.89) to allow the effective prediction of quadruplex formation within the original article's consensus (Figure 3).

Figure 2: PQS frequencies by their absolute differences between the original and the upgraded G4-iM Grinder's PQSfinder for 11359 PQS. The horizontal red lines give the accumulative percentage of PQS which difference is inside the range.

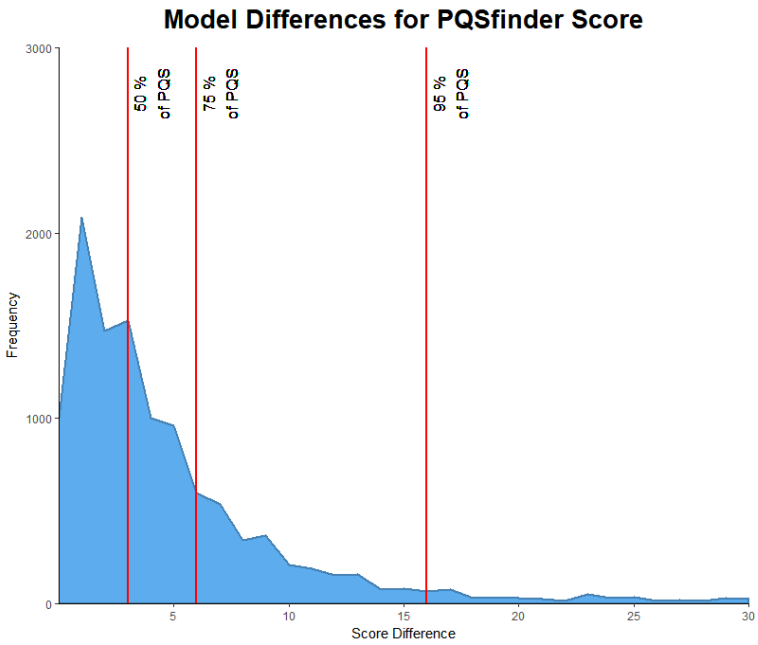
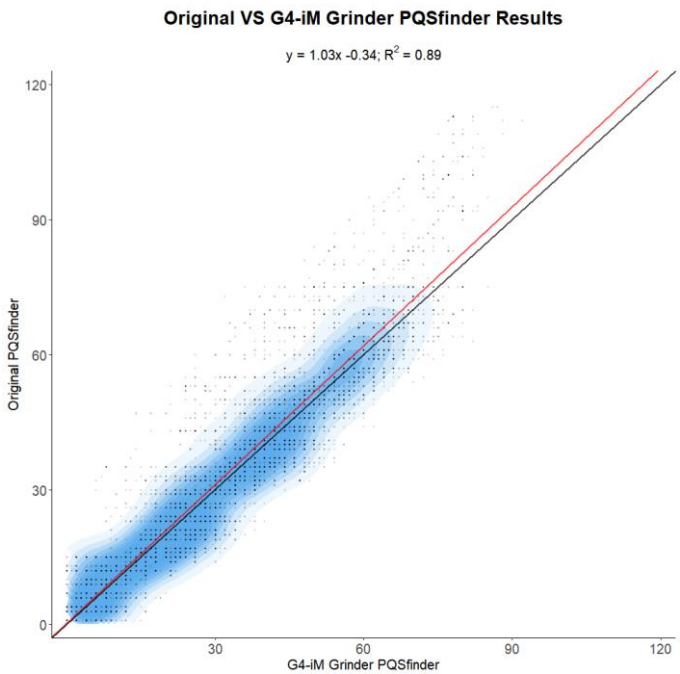


Figure 3: Relationship between the scoring of the original and G4-iM Grinder's PQSfinder systems for 11359 PQS. The red line is the found correlation line (y = 1.03x - 0.34), and in black the expected perfect match (y = 1x + 0). As a blue shadow, 2D PQS density between both scoring systems.



c) G4Hunter:

This scoring system was directly implemented from the original article (1). However, it was modified to fit into a -100 to 100 % score by multiplying the original score by 25 (the original scale ranges from -4 to 4).

d) Final Score Evaluation: *WeightedParameters* and *FreqWeight*.

After applying all the desired scoring systems, a final score value is calculated modulated by the *WeightedParameters* and *FreqWeight* variables.

WeightParameters: is a vector of 3 numbers that state the weight of each scoring system in the final score of a PQS. The first value within the variables is given to the G4Hunter score, the second to PQSfinder and the third to cGcC. Each part is dependent on the activation of its score. This means that if for example *G4Hunter* = FALSE, even if *WeightParameters* = (50, 50, 50), it would calculate the final score assuming the vector is (0, 50, 50).

This value is predefined to be (50, 50, 0) because cGcC has a different scaling system which is not compatible with G4Hunter and PQSfinder. Some cGcC scores can reach values of 2000 and therefore it is recommended to interpret cGcC separately.

*FreqWeigh*t: A constant that gives the importance of the sequence frequency. Useful only for Method 2B and 3B, where frequency of the structures are calculated.

The Final Score is calculated when at least two scoring systems are activated. It will use the results of each scoring system and apply it in a weighted mean formula, where the weight of each value is defined within the variable *WeightParameters*. For Method 2A and 3A, this is:

$$Score' = \frac{G4H.Score \times W.P[1]}{\sum W.P.} + \frac{PQSfinder.Score \times W.P[2]}{\sum W.P.} + \frac{cGcC.Score \times W.P[3]}{\sum W.P.}$$

Where *G4H.Score* is the Score of G4Hunter, *W.P.* is the *WeightParameter* variable

For Method2B or Method3B an appendix is added to the formula which takes into account the frequency of the structure using *FreqWeight* to modulate the importance of this characteristic.

$$Score'' = Score' + (FreqWeight \times log_{10}(Freq.))$$

Where *Freq.* is the frequency of the potential quadruplex.

e) Scoring system for i-Motifs

As G4Hunter punctuates negatively C presence, it was decided to apply an equal but contrary scale for i-Motif likeliness of forming to that of the G based PQSs. This means that when evaluating under the scoring system implemented here, more positive results will mean a higher probability of a PQS being an actual G4; and on the other hand, larger negative values will mean a higher probability of a potential i-Motif (PiMS) being an actual i-Motif. Hence for the analysis of PiMS (*RunComposition* = C), the cGcC formula is elevated and multiplied by -1 and the PQSfinder value is multiplied by -1 as to adapt them to this i-Motif scale.

G4Hunter and G4-iM Grinder's adaptation of PQSfinder were used to punctuate the *in vitro* probability of formation of 96 known-to-form i-Motif sequences. When applying these score systems to the verified i-Motifs, the results were:

G4Hunter = -49.5 ± 17.0 PQSfinder = -62.1 ± 10.7 Score(mean) = -55.8 ± 13.6

This confirms the direct relationship between high (absolute) probability formation scores and their actual *in vitro* formation (Figure 4, Top: boxplots). These results are very similar to those observed in G4s, which have been used by the original articles (and G4-iM Grinder) to verify the link between high scores and higher potential of quadruplex potential. Moreover, the available melting temperature (Tm) per confirmed sequences was also used to find any correlations between them, the length and the scores (Figure 5). pH 5 Tm and Score(mean) displayed the most interesting correlation (Figure 4, graphs). In all cases, higher scores correlate with higher Tm

values, despite the experimental differences for the Tm acquisition, which limits the conclusion extrapolation.

Figure 4 - Top: Distribution of i-Motifs regarding their scores. Botton: i-Motif relationship study between their pH-dependent melting temperatures and their PQSfinder score (A), G4Hunter (B) and mean score (C).

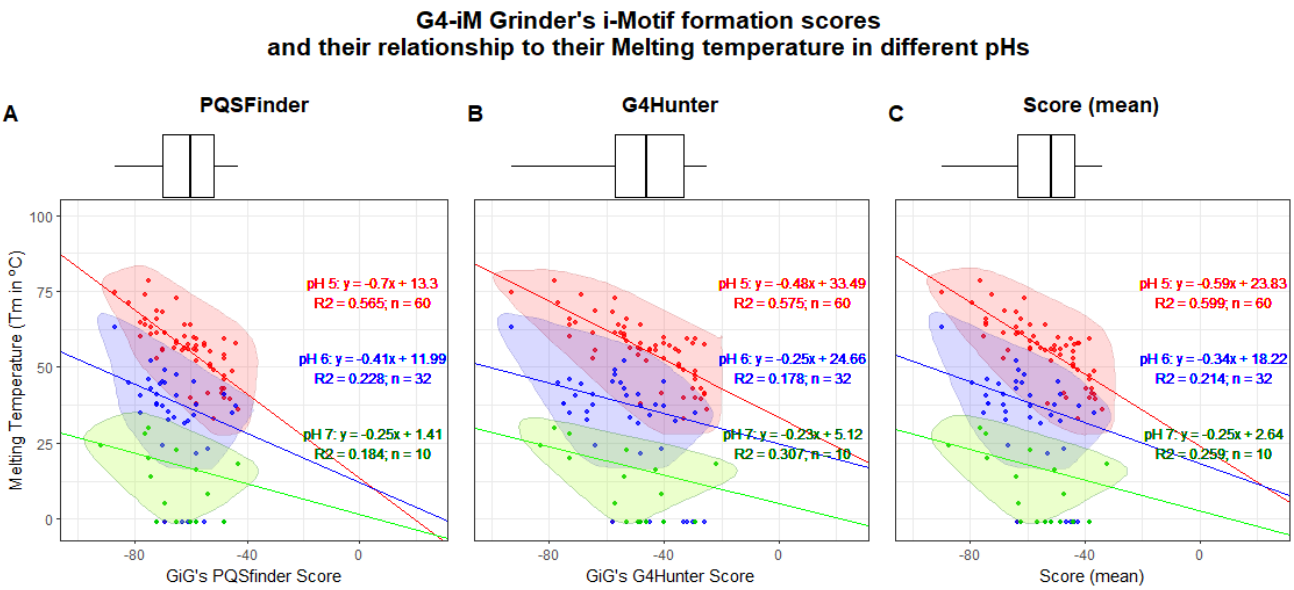
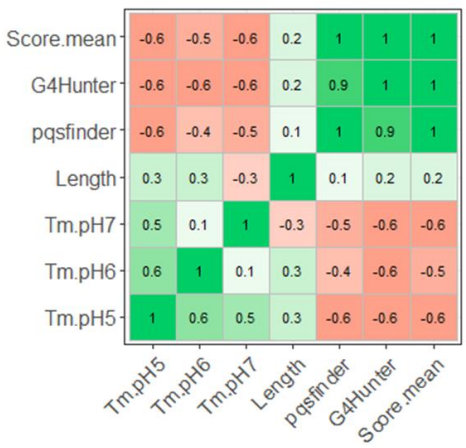


Figure 5: Correlation plot for several characteristics of the 96 i-Motif sequences examined.



6. Explanation of other Variables:

These variables can be further modified to adapt a G4-iM Grinder analysis.

Complementary: If TRUE, it will also create the complementary strand of the sequence and add it to the analysis. A new column will be added to state the sequence's strand location: original (+) or complementary (-).

DNA: Identifies whether the Sequence is a DNA sequence. If *DNA* = FALSE, it means the sequence is an RNA genome. Useful for sequence complementary conversion, *KnownQuadruplex* and *KnownNOTQuadruplex* functions.

These variables will qualify a G4-iM Grinder analysis in a score-independent manner.

LoopSeq: A vector which will create a new column per element within the variable to quantify the occurrence of the element in the sequence as a percentage.

For example, if *LoopSeq* = c(G, T, A, C, GGGTTA), it will create 5 columns in the result tables of Method 2 and 3 with the percentage of the final sequence (PQS/ i-motif) which is G, T, A, C and GGGTTA respectively.

KnownQuadruplex: A function that will detect *in vitro* confirmed quadruplex within G4-iM Grinder's Results. For PQS in DNA searches (*RunComposition* = "G", *DNA* = TRUE) most sequences were downloaded from G4Hunter's supplementary material (1). For PQS in RNA searches (*RunComposition* = "G", *DNA* = FALSE), sequences were downloaded from G4RNA DDBB (scottgroup.med.usherbrooke.ca/G4RNA) (4). Several known-to-form DNA i-Motifs (*RunComposition* = "C", *DNA* = TRUE) were also searched for and included. All these sequences were modified to start and end with the first and last G or C of the sequence as to mimic G4-iM Grinder's results. Duplicated sequences were eliminated. In total, this DDBB includes 294 DNA and 223 RNA sequences which form G4, and 100 DNA sequences that form i-Motifs. The name/s of the detected quadruplex structures for each sequence are stored in a new column. The number

of times detected is stated after the name in between brackets (the search is done in an overlapping manner).

KnownNOTQuadruplex: A function that will detect *in vitro* confirmed sequences that do NOT form Quadruplexes within G4-iM Grinder's Results. For PQS in DNA searches (*RunComposition* = "G", *DNA* = TRUE) sequences were downloaded from G4Hunter's supplementary material (1). For PQS in RNA searches (*RunComposition* = "G", *DNA* = FALSE), sequences were downloaded from G4RNA DDBB (scottgroup.med.usherbrooke.ca/G4RNA) (4). All these sequences were modified to start and end with the first and last G or C of the sequence as to mimic G4-iM Grinder's results. Duplicated sequences were eliminated as were sequences with less than 18 nucleotides. In total, the DDBB includes 89 DNA and 176 RNA which do NOT form G4. The name/s of the detected non-quadruplex sequences are stored in a new column. The number of times each NOT forming sequences is detected is stated after the name in between brackets (the search is done in an overlapping manner).

7. G4-iM-Grinder performance

To test the algorithm, human chromosome 22 (48.5 Mb) was loaded from the ensemble ftp server (version GCA_000001405.25) and analysed with G4-iM Grinder. Performance times and results of all the methods (M2A, M3A and M2B) were analysed with several variable configurations, to measure how these affect code execution (Table 5).

Table 5: Results of Chromosome 22 with different parameters. *BulgeSize* = 0, 1 or 2, *MinLoopsize* = 0 or 3, RunComposition = G, T, A or C, and *NCores* = 1 or 7. The rest of the variables were maintained fixed at predefined values, except for cGcC which was changed to TRUE as to evaluate it too and *Complementary* was set to FALSE to measure the differences between the number of structures found within runs of complementary bases. All the analyses were done using an Intel Core i7-4790 K CPU @ 4.00GHz with 16 Gb of RAM.

Variable configuration								
GrunCompositon	G	G	G	G	G	C	T	A
Ncores	1	1	1	1	7	7	7	7
BulgeSize	0	1	1	2	2	2	2	2
MinLoopSize	0	3	0	0	0	0	0	0
Performance (min)								
Sequence	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
Method 1A	0.07	0.16	0.13	0.16	0.14	0.14	0.15	0.17
Method 1B	0.00	0.03	0.02	0.05	0.05	0.05	0.06	0.06
Method 2A	1.35	4.47	5.73	9.91	9.15	8.80	9.98	11.66
Method 2B	0.00	0.01	0.03	0.03	0.03	0.03	0.06	0.06
Method 3A	1.85	7.38	11.25	25.07	3.57	3.57	4.50	4.83
Method 3B	0.00	0.01	0.03	0.03	0.03	0.03	0.06	0.06
G4Hunter	0.69	1.54	3.52	4.92	0.54	0.49	0.78	0.86
PQSfinder	0.00	0.00	0.00	0.00	0.00	0.00	*NA*	*NA*
cGcC	0.12	0.28	0.63	0.89	0.94	0.92	*NA*	*NA*
meanScore	0,00	0.00	0.01	0.00	0.00	0.00	0.00	0.00
Quantification	0,00	0.09	0.49	0.01	0.02	0.01	0.02	0.03
KnownQuadruplex	0.08	0.22	0.47	0.69	0.73	*NA*	*NA*	*NA*
Total	**4.17**	**14.20**	**22.32**	**41.77**	**15.21**	**14.05**	**15.61**	**17.74**
Results - Structures found								
Method 2A	17320	35041	97098	116290	116290	117428	210556	228603
Method 2B	15906	32216	87690	107945	107945	108154	175696	183499
Method 3A	8260	24760	39421	70858	70858	69714	96983	103060

Both performance time and results show dependency on the sequence analysed, its composition and organization, and the parameters employed. The simplest of options examined here (*BulgeSize* = *MinLoopsize* = 0) yielded 17320 PQS. This is 1.14 million PQS when extrapolated to the whole genome, a number similar to previous estimates for folding rule abiding-structures (5). These results increased fivefold when the number of acceptable bulges (*BulgeSize)* was increased to 1. When they were set to 2, smaller differences were observed because of the user-defined variable *MaxIL* that limits the maximum number of total acceptable bulges in the sequences.

These results were compared to those of other nucleotide-composition runs, including potential i-Motif sequences which are similar in number to G-based PQS. For structures composed of T and A runs, which have no known physical meaning, this count almost doubles that of G and C results. This is in accordance with a previous report by Huppert (5).

Regarding performance, the most time-consuming processes were M2A, M3A and G4Hunter scoring system. The total process time was optimized by increasing the numbers of cores of the computer to do the calculations. Using 7 cores reduced the time by 66 % in total. As a comparison, PQSfinder was downloaded directly from CRAN (December 2017) and executed with the same genomic sequence to compare the execution times. In total, the analysis with PQSfinder took 4 hours and 16 minutes to finish when running with predefined variables (except strand which was set to “+”).

8. Grinder Package

G4-iM Grinder can be download from github: EfresBR/G4iMGrinder. G4-iM Grinder requires the installation of other CRAN based and Bioconductor packages. Please, ensure all required packages are installed. G4-iM Grinder was successfully downloaded and tested in MacOS 10.12.6, Windows 10 (x64) and Ubuntu 18.04.1 (x64) running R 3.6.0 (and older) and R studio 1.2.1135 (and older). In Ubuntu, the installation of devtools may require further effort (link). Other OS including x86 systems have not been tested.

```
pck <- c("stringr", "stringi", "plyr", "seqinr", "stats", "parallel",
"doParallel", "beepr", "stats4", "devtools", "dplyr", "BiocManager")

#foo was written by Simon O'Hanlon Nov 8 2013.
#Thanks Simon, thanks StackOverflow and all its amazing community.

foo <- function(x){
  for( i in x ){
    #  require returns TRUE invisibly if it was able to load package
    if( ! require( i , character.only = TRUE ) ){
      #  If package was not able to be loaded then re-install
      install.packages( i , dependencies = TRUE )
      #  Load package after installing
      require( i , character.only = TRUE )
    }
  }
}
foo(pck)
BiocManager::install(c("BiocGenerics", "S4Vectors") , ask = FALSE, up-
date = TRUE)
```

To install and load G4-iM Grinder

```
devtools::install_github("EfresBR/G4iMGrinder")

library(G4iMGrinder)
```

Executing a genomic G-Quadruplex analysis with **G4iMGrinder** function

```
# Using a genome available online
loc <- url("http://tritrypdb.org/common/downloads/release-36/Lma-
jor/fasta/TriTrypDB-36_Lmajor_ESTs.fasta")

Sequence <- paste0(seqinr::read.fasta(file = loc, as.string = TRUE,
legacy.mode = TRUE, seqonly = TRUE, strip.desc = TRUE), collapse = "")

# Executing a grind on the sequence in search of PQS
Rs  <- G4iMGrinder(Name = "LmajorESTs", Sequence = Sequence)

# Forcing the folding rule to the limit (this will take longer)
Rs2 <- G4iMGrinder(Name = "LmajorESTs", Sequence = Sequence, BulgeSize
= 2,   MaxIL = 10, MaxLoopSize = 20)
```

Summarizing an analysis with **GiGList.Analysis** function to compare the results between genomes. This will quantify the number of results and density of each analysis. It will also give the number of results that have at least a minimum frequency, score and size. These variables can be modified. See the package documentation for more information regarding **GiGList.Analysis**.

```
# summarizing first search
ResultTable <- GiGList.Analysis(GiGList = Rs, iden = "Predefined")
# adding the second analysis in a new row
ResultTable[2,] <- GiGList.Analysis(GiGList = Rs2, iden= "ForceLimit")
```

Executing an analysis of a higher order structure with **GiG.M3Structure** to analyse its potential subunit configuration. This will give all and the most interesting subunit conformations as stated in the article. See the package documentation for more information regarding **GiG.M3Structure**.

```
# analyzing the PHOQS structure in row 1 of data frame PQSM3A in Rs.
PHOQS126 <- GiG.M3Structure(GiGList = Rs, M3ACandidate = 1, MAXite =
10000)
```

Finding the reference for the Known-To-Form Quadruplex structures of an interesting Result. This procedure is the same for Known-NOT-To-Form sequences.

```
# The PHOQS structure in row 126 has the known-to-form sequence 93del.
RefPHOQS126 <- Known_to_form_Quadruplex[
                        Known_to_form_Quadruplex$Name == "93del",
                        c("Ref", "DOI")]
```

Updating a G4-iM Grinder analysis with different variables using the **GiGList.Updater** function. This will avoid doing a new search analysis on the sequence and therefore will be more time and resource efficient.

```
# As the PHOQS structure in row 126 looks promising, we will also
# examine the Known-NOT-to-form Quadruplex of the results,
# quantify the % of GGG and TTA present in the sequence, and modify
# the score and frequency weight of the final score.
Rs3 <- GiGList.Updater( GiGList = Rs,
                    KnownNOTQuadruplex = TRUE, KnownQuadruplex = TRUE,
                    LoopSeq = c("GGG", "TTA"),
                    FreqWeight = 100, WeightParameters = c(75, 25, 0))
```

To search for potential i-Motifs in the genome we can repeat the analysis with **G4iMGrinder** function changing *RunComposition* = "C". However, if a previous analysis of the genome has already been done with the complementary base-pair, we can also use the function **GiGList.Updater** to search for the resulting opposite structures.

```
# Doing a grind in search for i-Motifs in the sequence

Rs_iM1 <- G4iMGrinder(Name = "LmajorESTs", Sequence = Sequence,
RunComposition = "C")

# Using the previous PQS search to locate the base-pair PiMS.
# This is more efficient. However as the sequences change, the
# previous analysis will be deleted. If you desire to vary the
# analysis with non-predefined values, they should be stated here.

Rs_iM2 <- GiGList.Updater(GiGList = Rs, ChangeRunComposition = TRUE)
```

9. Genomes used and results with all methods

The complete human chromosomic genome (GRCh38.p12, Genome Reference Consortium Human Build 38, INSDC Assembly GCA_000001405.27– gh38) was downloaded from the Sanger Institute (www.sanger.ac.uk) in May 2019. Other non-human genomes used in this paper were downloaded in May 2019 from https://www.ncbi.nlm.nih.gov/, and include:

Name	iden	RefSeq
Aspergillus fumigatus	Fungus	NC_007194.1 to NC_007201.1
Candida albicans	Fungus	NC_032089.1 to NC_032096.1
Candida auris	Fungus	PGLS00000000.1
Candida parapsilosis	Fungus	HE605209.1
Pneumocystis jirovecii	Fungus	LFWA00000000.1
Saccharomyces cerevisiae	Fungus	NC_001133.9 to NC_001224.1
Spizellomyces punctatus	Fungus	ACOE00000000.1
Brucella melitensis	Gram:Negative	NC_003317.1 to NC_003318.1
Escherichia coli	Gram:Negative	NC_011750.1
Klebsiella pneumoniae	Gram:Negative	NC_016845.1
Legionella pneumophila	Gram:Negative	NC_002942.5
Neisseria meningitidis	Gram:Negative	NC_003112.2
Pseudomonas aeruginosa	Gram:Negative	NC_017548.1
Salmonella enterica	Gram:Negative	NC_003197.2
Treponema pallidum	Gram:Negative	NZ_CP003679.1
Vibrio cholerae	Gram:Negative	NC_002505.1 to NC_002506.1
Yersinia pestis	Gram:Negative	NC_003143.1
Bacillus anthracis	Gram:Positive	NC_003997.3
Clostridium botulinum	Gram:Positive	NC_009698.1
Clostridium tetani	Gram:Positive	NC_004557.1
Corynebacterium diphtheriae	Gram:Positive	NZ_LN831026.1
Mycobacterium leprae	Gram:Positive	NC_002677.1
Mycobacterium tuberculosis	Gram:Positive	NC_000962.3
Streptococcus pneumoniae	Gram:Positive	NC_003098.1
Streptococcus pyogenes	Gram:Positive	NC_002737.2
Echinococcus granulosus	Helminth	APAU00000000.2
Echinococcus multilocularis	Helminth	CBLO000000000.2
Schistosoma mansoni	Helminth	NC_031495.1 to NC_002545.1
Trichuris trichiura	Helminth	CBXK000000000.1
Entamoeba histolytica	Protozoan	AAFB00000000.2
Leishmania donovani	Protozoan	NC_018228.1 to NC_018263.1
Leishmania infantum	Protozoan	NC_009386.2 to NC_009420.2
Leishmania major	Protozoan	NC_001905.3 to NC_007287.2
Plasmodium falciparum	Protozoan	NC_004325.2 to NC_037283.1

Toxoplasma gondii	Protozoan	NC_031467.1 to NC_001799.1
Trypanosoma brucei	Protozoan	NC_026734.1 to NC_026744.1
Trypanosoma cruzi	Protozoan	NZ_AAHK00000000.1
Dengue virus	Virus	NC_001474.2
EpsteinBarr virus	Virus	LS992244.1
Hepacivirus C	Virus	NC_004102.1
HIV-1	Virus	NC_001802.1
Influenza A virus	Virus	NC_007366.1
Measles morbillivirus	Virus	NC_001498.1
Polio virus	Virus	NC_002058.3
Rabies lyssavirus	Virus	KY982923
Rubella virus	Virus	NC_001545.2
Variola virus	Virus	NC_001611.1
Zaire ebolavirus	Virus	NC_002549.1
Zika virus	Virus	NC_012532.1

The complete G4iM-Grinder Results used for this article can be downloaded from github: EfresBR/G4iMGrinder (https://github.com/EfresBR/G4iMGrinder). These results are compressed in the archive named "Results.RAR". Inside, four RData images are included: "*hg39PQS.RData*" (Human G-based PQS), "*hg39PiMS.RData*" (Human C-based PiMS), "NonHumanPQS.RData" (non-human G-based PQS) and "NonHumanPiMS.RData" (non-human C-based PiMS).

For human results, 25 G4-iM Grinder Lists (GiG-Lists) and 1 data frame exist within each Rdata archive. 24 of these lists are the results for each individual human chromosome and the last (and biggest) is the complete and joint M2B and M3B evaluation of the genome (it is **NOT** the sum of the individual chromosomes). The data frame *ResultTable* is a summary of the results, after using G4-iM Grinder's Analysis function which helps extract the basic information of the results in a fast and convenient way.

For non-human results, each GiG-List is the analysis of an organisms. The data frame *A.ResultsTable* is a summary of the results, after using G4-iM Grinder's Analysis function. *A.GenomeTable* contains the information of each genome examined.

10. References

1. Bedrat,A., Lacroix,L. and Mergny,J.-L. (2016) Re-evaluation of G-quadruplex propensity with G4Hunter. *Nucleic Acids Research*, **44**, 1746–1759.

2. Hon,J., Martínek,T., Zendulka,J. and Lexa,M. (2017) pqsfinder: an exhaustive and imperfection-tolerant search tool for potential quadruplex-forming sequences in R. *Bioinformatics*, **33**, 3373–3379.

3. Beaudoin,J.-D., Jodoin,R. and Perreault,J.-P. (2014) New scoring system to identify RNA G-quadruplex folding. *Nucleic Acids Research*, **42**, 1209–1223.

4. Garant,J.-M., Luce,M.J., Scott,M.S. and Perreault,J.-P. (2015) G4RNA: an RNA G-quadruplex database. *Database*, **2015**.

5. Huppert,J.L. (2005) Prevalence of quadruplexes in the human genome. *Nucleic Acids Research*, **33**, 2908–2916.