G4-iM Grinder: DNA and RNA G-Quadruplex, i-Motif and higher order structure search and analyser tool

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

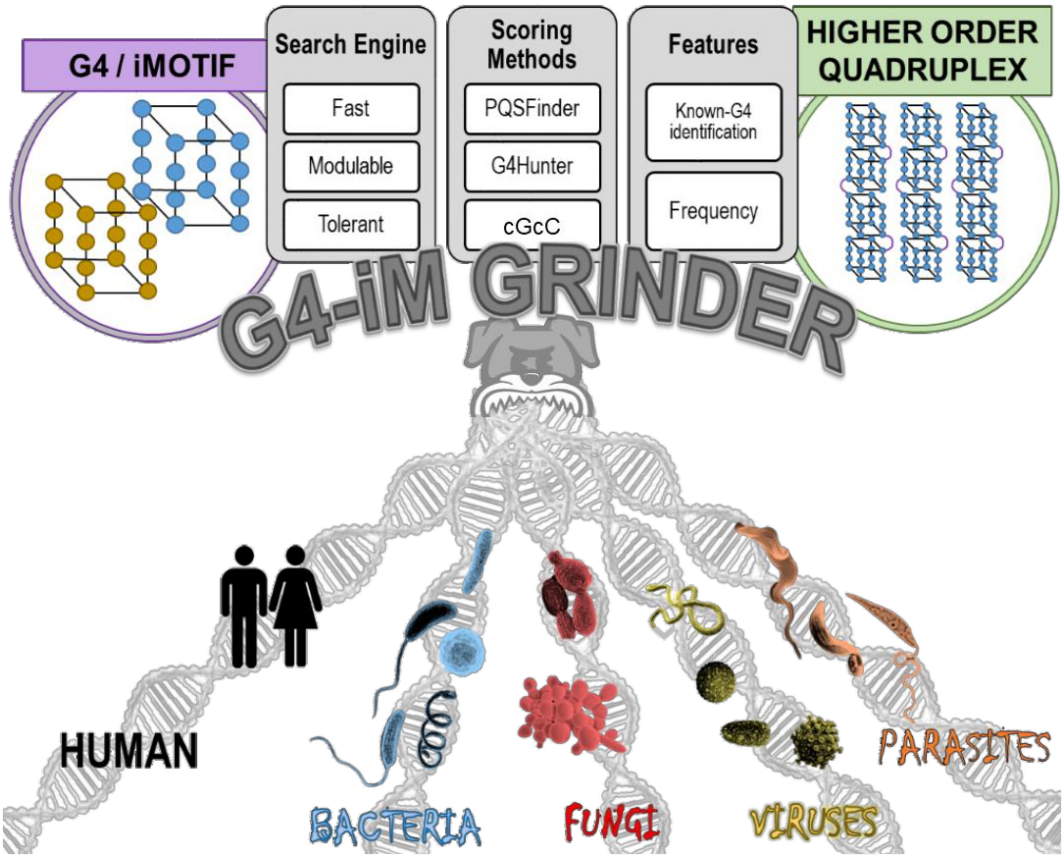
[1] Department of Biochemistry and Molecular Pharmacology, Instituto de Parasitología y Biomedicina López Neyra, CSIC, PTS Granada, Avda. del Conocimiento, 17, 18016 Armilla, Granada, Spain.

[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

Herein, we present G4-iM Grinder as a system for potential G4, i-Motif and higher-order structure identification and characterization. Several grading tools of biological relevance and G4 *in vitro* formation probability are included in this highly modulable and robust engine. G4-iM Grinder improves other current quadruplex search engines when compared in capabilities and processing time. We used G4-iM Grinder in the analysis of the complete human genome whilst focusing on frequency and score of G-based structures. We studied the most recurrent potential quadruplex sequences (PQS) and the longest highest scoring potential higher order quadruplex sequences (PHOQS) in our genome. As proof of the analytical capabilities of G4-iM Grinder, we also analysed a new PHOQS and predicted the most probable PQS subunits to form it. Taking the human average PQS density as reference, we examined the genomes of organisms that cause leishmaniosis, diphtheria, brucellosis, meningitis, pneumonia, toxoplasmosis, tuberculosis, leprosy, AIDS, dengue fever and hepatitis C, and found they are very rich in PQS. G4-iM Grinder identified within many of these organisms several already known-to-form G4 sequences. Together, this suggests that G4-quadruplexes may potentially be important therapeutic targets against many of 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 abundantly studied in the last 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})$_4$ 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 literally follow the folding rule. Loop range inconformity, 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 like hood of G4 formation for each result is here defined 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 is able to detect a fix 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 upmost 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 G4Hunter(15) and G4 potential calculator(16) use this statistical analysis window that willingly not defines individual

PQS boundaries nor defect types. Hence, they are able to 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 trained on over 200 structures to classify folding rule abiding PQSs.(5)

All quadruplex search models have several drawbacks and limitations despite the advances in the field. For most, 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, none of the current search engines takes into account nor calculates genomic PQS frequency from the results. Even if 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, it was 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 also exist.(26)

To finish, 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 then can assemble 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

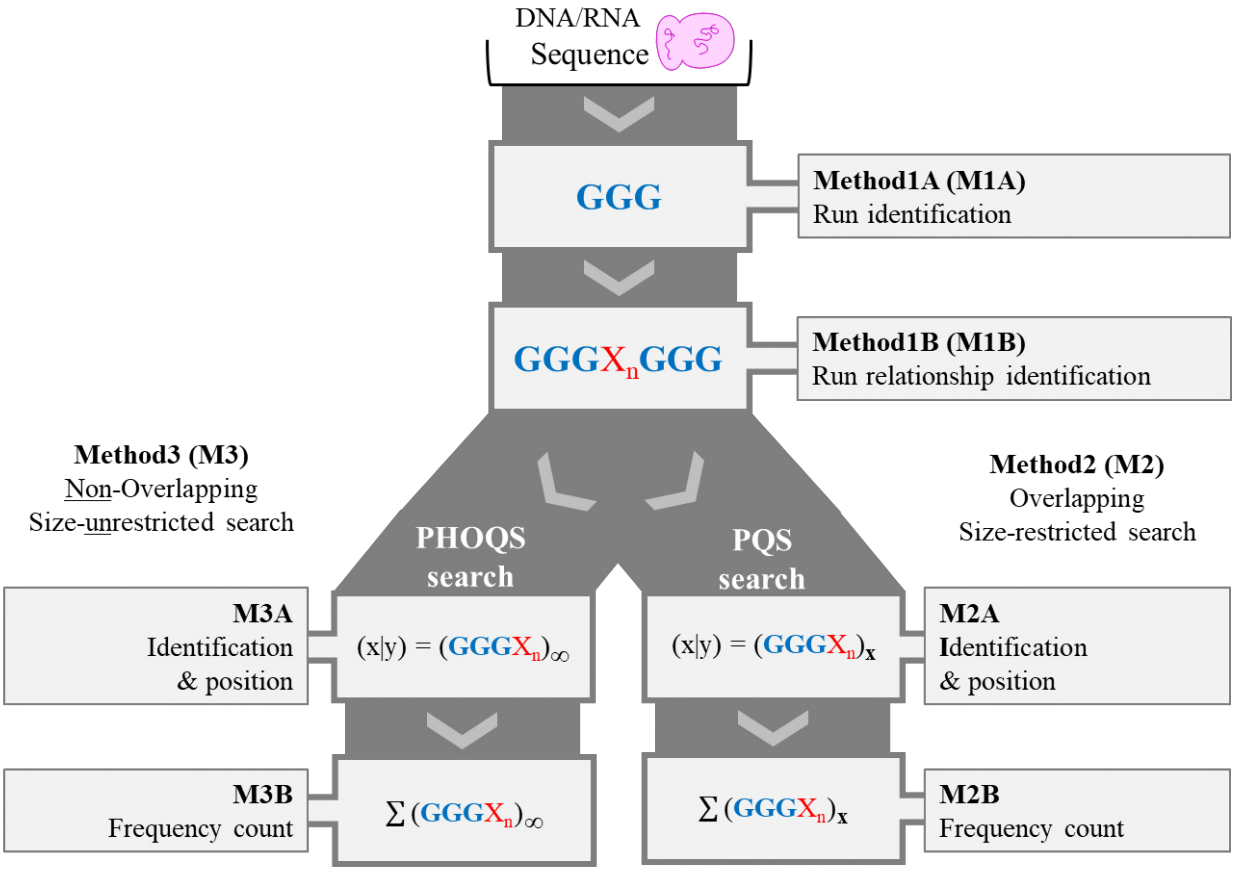
G4-iM Grinder

Our contribution to the field is focused on solving these limitations in an easy and fast manner for the user. The solution was developed with several main separate objectives, the first of which was to allow extensive user customization and liberty to the search parameters. These were then to be introduced to a fast, reliable and tolerant search engine capable of detecting even potential higher-order structures. The second objective was to expand the analyses options of the search engine output. The result is G4-iM Grinder, a search engine and analyser with 13 modulable variables, three analytical methods, five possible scoring systems and several quantification and qualification tools with the ability to work in parallel in various processors (Supplementary information – S1, Classification performance). Structure frequency analysis was included as an alternative or a compliment to other scoring systems such as G4Hunter, PQSfinder and cGcC. PQSfinder was adapted and upgraded using machine learning to allow potential higher order structures and irregular PQS punctuation. Optionally and when looking for PQSs, G4-iM Grinder's results can be analyzed to detect any already verified G4-quadruplexes within a list of confirmed known-to-form sequences.

Algorithm

The algorithm code is written in R and divided into several parts. Initially a setup is executed and 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, being the first two the search engine (Method 1A and 1B, M1A and M1B respectively, Figure 1), and the other two the analysers (Method 2 for overlapping and size-dependent search, and Method 3 for non-overlapping size-independent examination, M2A and M3A respectively).

Figure 1. G4-iM Grinder algorithm pathway when *RunComposition* = G to find potential quadruplex sequences (PQS) and potential higher order quadruplex sequences (PHOQS).



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 – S2. 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 runs that have no link with the previous yet are linked with the posterior 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 identified all results with M2A and M3A, 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.

When each analysing method finishes, the outcome will get examined by *LoopSeq* function (to quantify the % of the structure which is a predefined pattern) and *KnownG4* function (to detect sequences which have already been demonstrated to form *in-vitro* G4 structures, Supplementary information – S.3. Other variables). Then, each structure can be evaluated by G4Hunter as a scoring method to determine the *in vitro* potential of formation. PQSfinder and cGcC scoring methods have been adapted from the original articles through machine learning and are also available to use (Supplementary information – S.4. Scoring). If several scoring systems are selected, a final score will be calculated using a weighted average formula modulated by the

variable *WeightedParameters*. The predefined values of G4-iM Grinder's final score is set to be the average between G4Hunter and PQSfinder scores for DNA PQS. For M2B and M3B, this final score will be calculated taking into account the structure frequency as well. Additionally and just for i-Motifs, all the scoring systems are modified to evaluate potential by an equal but contrary scale to G4, meaning that the best results for C-based structures are near or less than -100 whilst for G4s best results are near or over 100.

If the computer can and 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 execution.

Table 2. Modulable variables with explanation and predefined values for the G4-iM Grinder engine. An additional nine parameters can be modified for PQSfinder analysis (Supplementary information - S.4. Scoring).

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	Analyze the complementary strand of the imputed sequence.
	DNA	TRUE	As to tell G4-iM Grinder if the sequence is DNA. If false, it will assume its RNA. Useful for complementary conversion and *KnownG4*.
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	*KnownG4*	TRUE	Identification of known forming G4s (DNA or RNA). Only if *RunComposition* = G.
Scoring	*G4Hunter*	TRUE	To apply G4-hunter evaluation system to the detected final sequences.(15) 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.(4) It is dependent on *RunComposition* = G or C.
	cGcC	FALSE	To apply the cGcC system adopted from Beaudoin *et al.*(31) 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	10	Weight of the structure frequency in the final score value for scoring calculations of M2B and M3B.

Testing G4/iM-Grinder results and 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, as to measure how these affect code execution (Table 3).

Table 3. 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
KnownG4	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.(9) 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.(9)

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 time dropped 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 h and 16 min to finish when running with predefined variables (except strand which was set to "+").

RESULTS

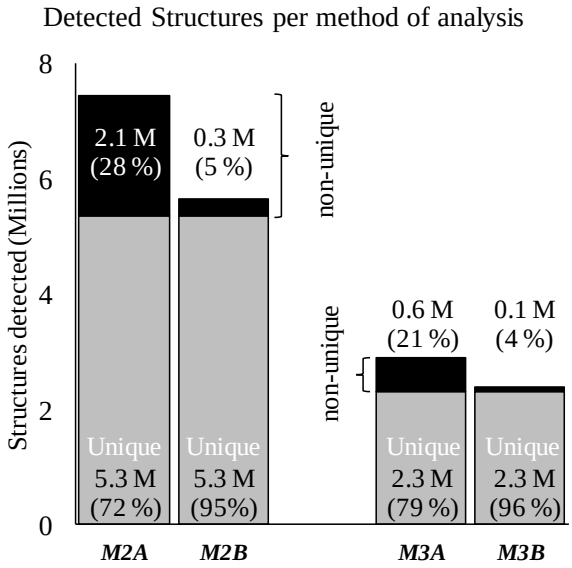
Full genomic analysis with G4-iM Grinder

A full analysis of all the human chromosomes was carried out with predefined variable configuration for the identification of potential G-forming PQS and C-based i-Motifs. Due to the small difference in results between using different *BulgeSize* values (as it is constrained by

MaxIL), it was decided to accept only 1 different nucleotide per run, for a maximum total of 3 per structure. It is well established that tetrad bulges are a factor for overall structure instability(32) and hence allowing too many of them would result in an increase in the detection rate of low probability *in vitro*-forming PQS plus an increase in the computation processing time.

The analysis (M2 and M3, A and B) of both strands took 15.75 h for G-based PQS and 13.78 h for C-based potential i-Motifs using 7 cores of a i7-4790 K CPU @ 4.00GHz. The same number of results were found for both G and C-based structures as the search included the complementary strand and these nucleotides are base pairs (Figure 2).

Figure 2. Detected structures in the entire human genome per method of analysis. The cumulative bars are divided in 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 in regard to the total is also showed in between parenthesis. Millions is abbreviated M.



Over 7.4 million PQS were detected with M2A (overlapping size-restricted method), and 2.9 million with M3A (non-overlapping size-unrestricted method). Non-unique sequences (% of PQS with a frequency of occurrence of over 1) represented 28 and 21% of all results respectively. Each of these non-unique PQS is repeated 6 times in average, yet some sequences exceed repetitions of over 10000 (Table 4). An example is -Table 4: M2B, entry 1- which is repeated 33642 times. This very recurrent PQS is probably part of a bigger G4 sequence as it can be combined with -M2B, entry 2, 3 and 4- to yield -M3B, entries 2 and 3. This suggests a possible relationship between all these PQS, such as the potential existence of a fluctuating PQS or being a part of a potential higher order structure. Another potential link can be appreciated between the first and fourth most frequent sequences found with M3B -M3B, entry 1 and 4- which differ in a single nucleotide in the 18th position (C to T).

Table 4. Most frequent PQS found in the complete human genome with methods M2B and M3B. In blue the G-runs, in green the bulges, in red the loops. Abbreviations: Freq. is frequency, IL is total bulges, G4H is G4Hunter score, PF is Pqsfinder score, % G is PQS composition that is G.

M2B - overlapping size-restricted method

	Freq.	Runs	IL	PQS	Length	G4H	PF	Score	% G
1	33642	4	3	GGGAGGCTGAGGCAGGAGAATGGCG	25	25	3	**59**	56
2	24502	4	3	GAGGCAGGAGAATGGCGTGAACCCGGG	27	13	1	**52**	48
3	15775	4	3	GGAGAATGGCGTGAACCCGGGAGGCGG	27	16	6	**56**	52
4	14351	5	4	GAGGCAGGAGAATGGCGTGAACCCGGGAGGCGG	33	16	2	**54**	52
5	13572	4	0	GGGAGGGAGGGAGGG	15	60	64	**107**	80

M3B - non-overlapping size-unrestricted method

	Freq.	Runs	IL	PQS	Length	G4H	PF	Score	% G
1	8766	5	3	GGGAGGCCGAGGCGGGCGGATCACGAGGTCAGGAG	35	23	11	**56**	54
2	4845	6	4	GGGAGGCTGAGGCAGGAGAATGGCGTGAACCCGGGAGGCGG	41	20	11	**54**	54
3	4639	6	4	GGGAGGCTGAGGCAGGAGAATGGCGTGAACCCGGGAGGCGGAG	43	20	6	**54**	53
4	3793	5	3	GGGAGGCCGAGGCGGGTGGATCACGAGGTCAGGAG	35	23	11	**54**	54

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 5). 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 5. Longest higher-order PQS (PHOQS) found in the human genome, which scores at least 50 with method 3A (M3A). G4Name are identified G4 sequences within the found structure which are known to form *in vitro*, followed by the times detected in between parenthesis. Abbreviations: Chrom. is chromosome, IL is bulges, G4H is G4Hunter score, PF is Pqsfinder score, G, C, A and T is PQS composition which is that nucleotide (as % of total).

Entry	Chrom.	Start	Length	Runs	IL	Strand	G4H	PF	Score	G	C	A	T	G4Name
1	11	400747	2343	305	13	-	48	68	58	66	17	7	10	
2	X	1587610	1666	293	37	-	60	62	61	70	1	29	1	
3	20	64093767	1646	151	13	+	44	62	53	60	8	13	19	
4	2	240923448	1351	156	0	-	42	65	54	59	9	7	25	
5	X	328170	1240	182	16	-	65	70	68	74	5	21	1	
6	10	131041965	1170	148	2	+	47	65	56	55	2	22	21	
7	7	470172	1125	136	0	-	44	65	54	58	8	17	16	
8	8	141812709	1102	142	8	-	44	63	54	60	9	13	18	
9	9	133668264	1089	189	11	-	45	57	51	66	1	27	6	
10	X	156029890	1005	164	6	+	44	59	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	-	51	68	60	63	19	4	14	d(G4C2)4 (3)
12	3	10005	978	159	1	-	43	61	52	55	2	15	29	TSG24 (44) X3ACT (3) Tet22 (2) G4CT-pallidum (6)

Potential higher order quadruplex sequence (PHOQS) analysis

Attention was set on HoEBR1, a relative 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 6). The known-to-form G4 sequence IV-1242540 was also located within these potential subunits.(33)

Table 6. 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. In blue the G-runs, in green the Bulges and in red the Loops of the subunits. G4Name are identified G4 sequences within the found structure which are known to form G4 *in vitro*, followed by the times detected in between parenthesis. Abbreviations: IL is total bulges in sequence, G4H is G4Hunter Score, PF is Pqsfinder score, G, C, A and T is PQS composition which is that nucleotide (as %).

Tag	Size	Runs	IL	PQS	G4H	PF	Score	G	C	A	T	G4Name
HoEBR1	118	17	3	GGGTCTGGGGAAAGAAGAAGAGGAGGAGGAGGAGGG GTTGTCGGGGGAAGAGGAGGAAAGGGAAGGGAATGA AGGGGGGAAGGGGAGGGGAAGGGGAGGGGGAGGGG	53	58	**62**	64	2	29	5	
Possible PQS subunits												
I	29	4	2	GGGTCTGGGGAAAGAAGAAGAGGAGGAGG	34	25	**30**	55	3	35	7	
II	31	4	2	GGGGAAAGAAGAAGAGGAGGAGGAGGAGGGG	41	28	**34**	61	0	39	0	
III	28	4	2	GAGGAGGAGGAGGAGGGGTTGTCGGGGG	47	35	**41**	68	4	18	11	
IV	26	4	2	GGAGGAGGAGGAGGGGTTGTCGGGGG	50	42	**46**	69	4	15	12	
V	32	5	3	GGAGGAGGAGGAGGGGTTGTCGGGGGAAGAGG	45	31	**38**	66	3	22	9	
VI	25	4	2	GAGGAGGAGGAGGGGTTGTCGGGGG	49	38	**44**	68	4	16	12	
VII	31	5	3	GAGGAGGAGGAGGGGTTGTCGGGGGAAGAGG	44	28	**36**	65	3	23	10	
VIII	29	4	2	GGAGGAGGAGGGGTTGTCGGGGGAAGAGG	46	34	**40**	66	3	21	10	
IX	28	4	2	GAGGAGGAGGGGTTGTCGGGGGAAGAGG	45	30	**38**	64	4	21	11	
X	26	4	2	GGAGGAGGGGTTGTCGGGGGAAGAGG	47	37	**42**	65	4	19	12	
XI	25	4	2	GAGGAGGGGTTGTCGGGGGAAGAGG	46	34	**40**	64	4	20	12	
XII	29	4	1	GGGGTTGTCGGGGGAAGAGGAGGAAAGGG	47	46	**46**	62	3	24	10	
XIII	25	4	1	GGGGGAAGAGGAGGAAAGGGAAGGG	47	46	**46**	64	0	36	0	
XIV	29	4	1	GAGGAGGAAAGGGAAGGGAATGAAGGGGG	41	42	**42**	59	0	38	3	
XV	27	4	1	GGAGGAAAGGGAAGGGAATGAAGGGGG	44	48	**46**	59	0	37	4	
XVI	26	4	1	GAGGAAAGGGAAGGGAATGAAGGGGG	42	45	**44**	58	0	39	4	
XVII	33	5	1	GAGGAAAGGGAAGGGAATGAAGGGGGGAAGGGG	48	50	**49**	61	0	36	3	
XVIII	26	4	0	GGGAAGGGAATGAAGGGGGGAAGGGG	57	66	**62**	65	0	31	4	
XIX	31	5	0	GGGAAGGGAATGAAGGGGGGAAGGGGAGGGG	60	69	**64**	68	0	29	3	
XX	26	4	0	GGGAATGAAGGGGGGAAGGGGAGGGG	63	71	**67**	69	0	28	3	
XXI	32	5	0	GGGAATGAAGGGGGGAAGGGGAGGGGAAGGGG	64	72	**68**	69	0	27	4	
XXII	23	4	0	GGGGGGAAGGGGAGGGGAAGGGG	78	78	**78**	78	0	22	0	
XXIII	29	5	0	GGGGGGAAGGGGAGGGGAAGGGGAGGGGG	79	81	**80**	79	0	21	0	IV-1242540 (1)
XXIV	21	4	0	GGGGAGGGGAAGGGGAGGGGG	81	80	**80**	81	0	19	0	IV-1242540 (1)
XXV	26	5	0	GGGGAGGGGAAGGGGAGGGGGAGGGG	81	80	**80**	81	0	19	0	IV-1242540 (1)
XXVI	32	6	0	GGGGAGGGGAAGGGGAGGGGGAGGGGAGGGGG	81	82	**82**	81	0	19	0	IV-1242540 (1)
XXVII	21	4	0	GGGGAAGGGGAGGGGGAGGGG	81	80	**80**	81	0	19	0	
XXVIII	27	5	0	GGGGAAGGGGAGGGGGAGGGGAGGGGG	81	83	**82**	82	0	19	0	
XXIX	32	6	0	GGGGAAGGGGAGGGGGAGGGGAGGGGGAGGGG	81	82	**82**	81	0	19	0	
XXX	21	4	0	GGGGAGGGGGAGGGGAGGGGG	86	85	**86**	86	0	14	0	
XXXI	26	5	0	GGGGAGGGGGAGGGGAGGGGGAGGGG	85	84	**84**	85	0	15	0	
XXXII	21	4	0	GGGGGAGGGGAGGGGGAGGGG	86	85	**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 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 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).

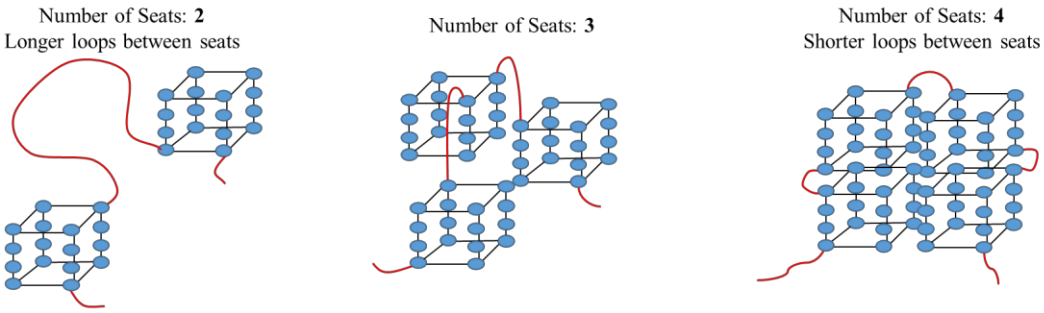
An alternative method based on randomly assigning *seats* to candidates was also developed and used. After 10000 iterations, the process identified all possible 307 subunit conformations that can give rise to HoEBR1. This was repeated ten times to make sure no conformations had been excluded. The 307 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 what 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 result 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*.

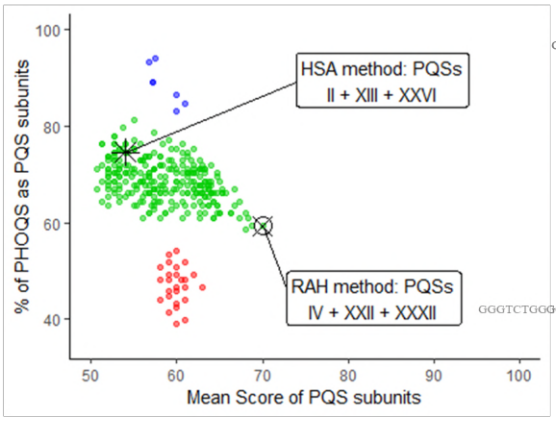
Figure 3. **A.** The PHOQS (potential higher order quadruplex sequence) HoEBR1 can be arranged into up to four seats. Greater number of *seats* means smaller loops between units and potentially a gain in structure stability. **B. 1.** High-scoring Sequential Allocation (HSA) conformation is based on 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 % 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 4 *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.

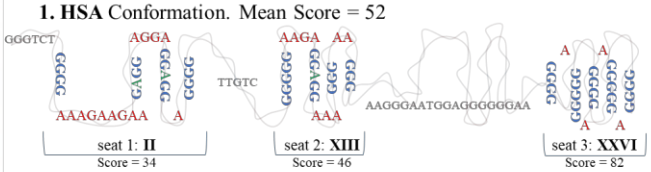
A.



B. Conformations of HoEBR1 arranged by their mean PQS score

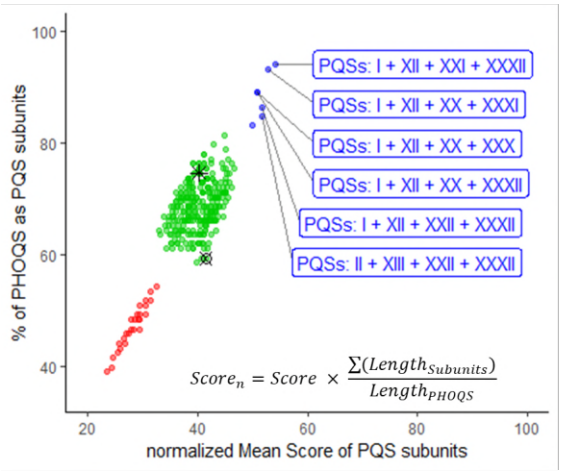


1. HSA Conformation. Mean Score = 52

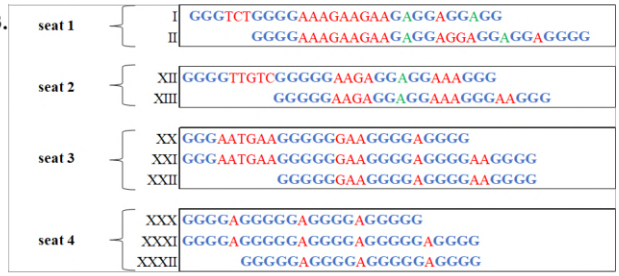


2. RAH Conformation. Mean Score = 70

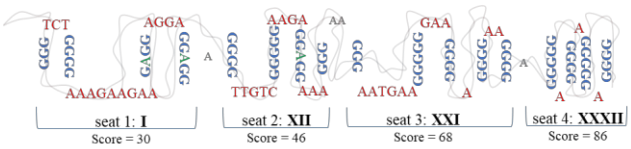
C. Conformations of HoEBR1 arranged by their normalized mean PQS score



3.



4. RAnH Conformation. Mean normalized Score = 54



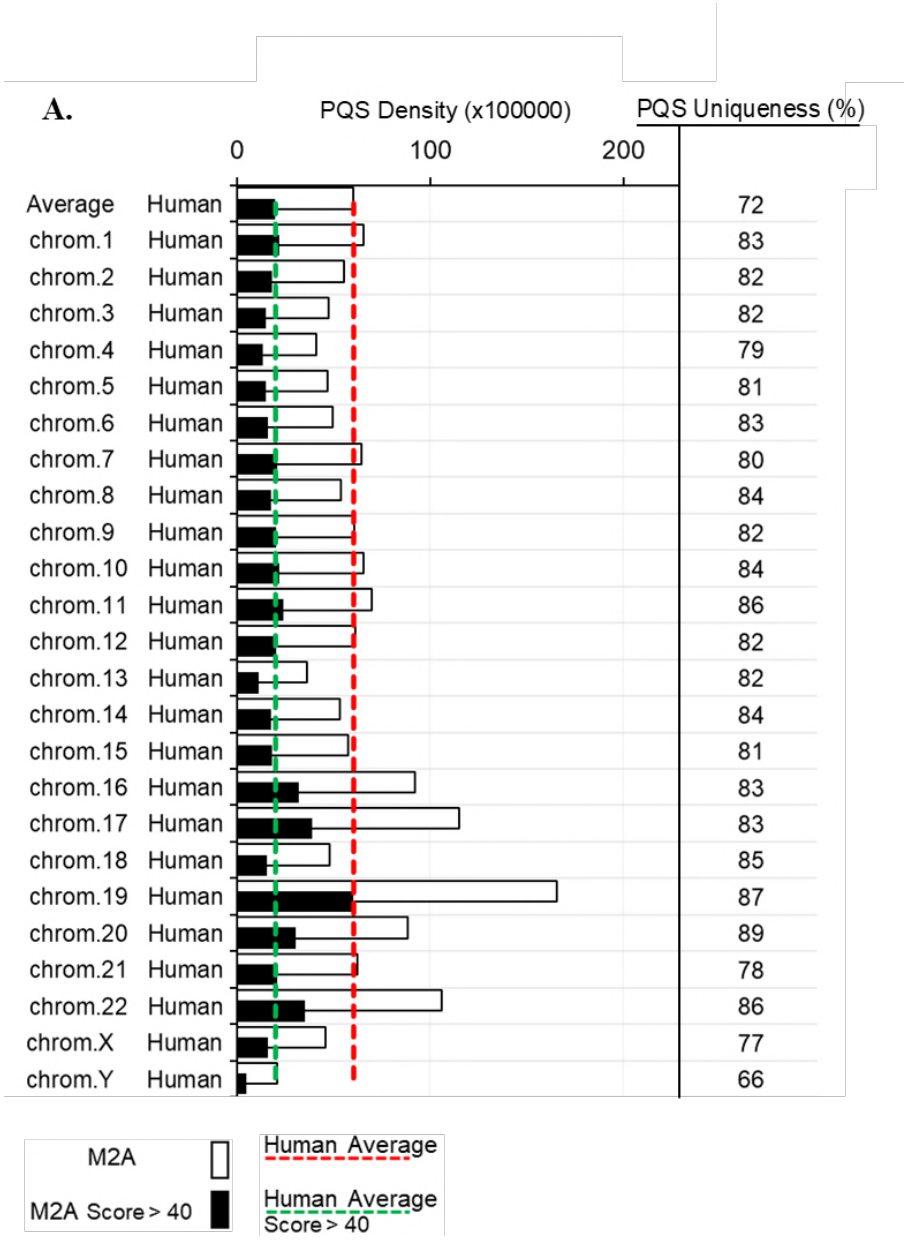
Potential G4 relevance in the genome of humans and other organisms

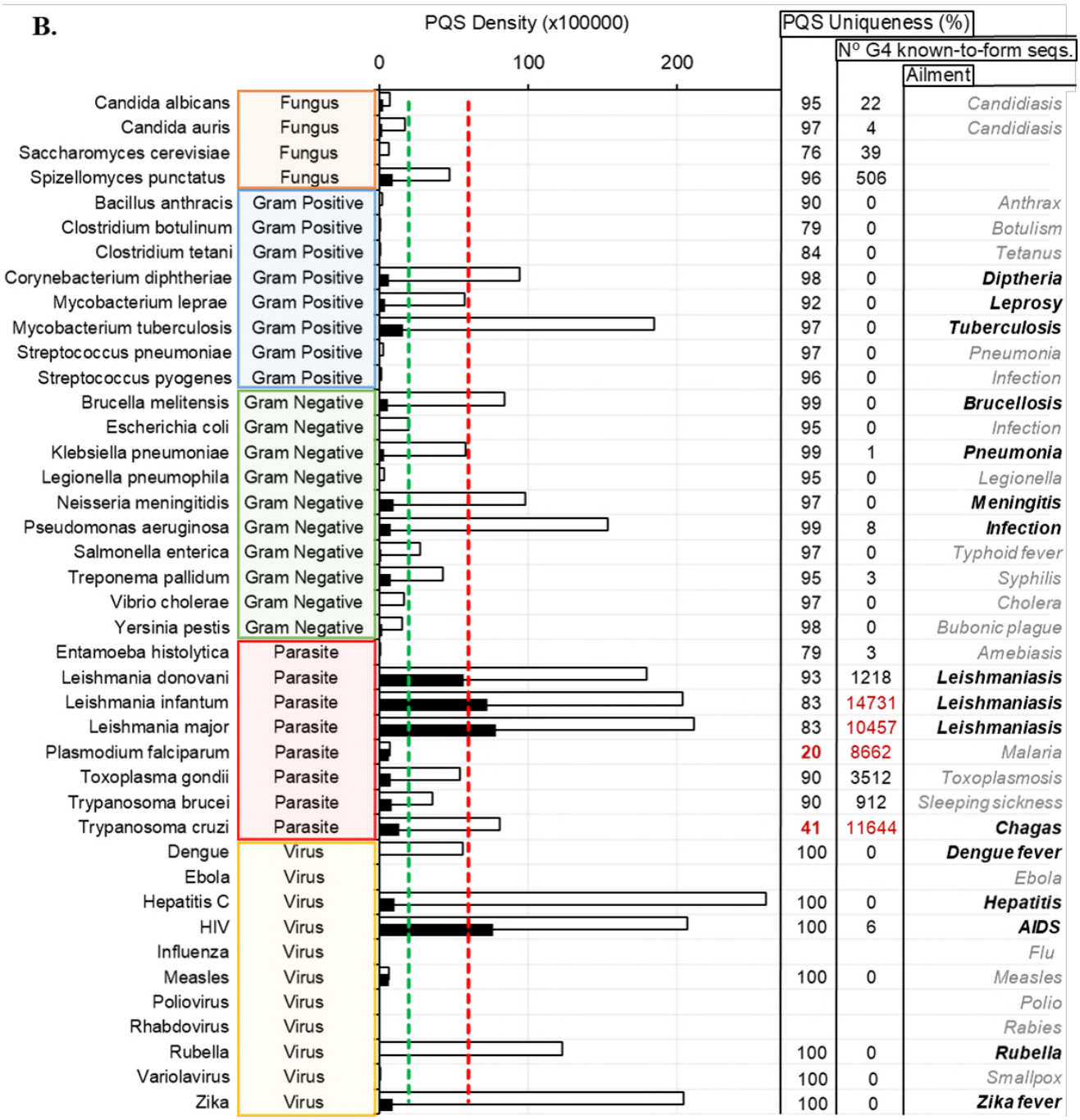
G4-iM Grinder results of the whole human genome were used to calculate the PQS density (per 100000 nucleotides) and % of non-unique sequences of each chromosome (Figure 4, A). PQS densities for structures with a minimum acceptable score (Score > 40) were also calculated and used to quantify the minimum most probable G4 density per chromosome. This threshold was obtained by applying mean minus standard deviation (54 - 13) of all the structure scores in the human genome which are already known to form G4 *in vitro*. The total human average density values were also calculated and then used as reference for the search results in other living organisms, including several: viruses, fungi, bacteria (both Gram positive and Gram negative) and protozoan parasites. This combination of applying various scoring criteria, analytical methods and genomic sequences allowed a wider context of interpretation.

Depending on the G4-iM Grinder method employed and the scoring criteria used, the human genome PQS density oscillates between 10 and 200 PQS per 100000 nucleotides. Chromosome 19 showed the highest PQS 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, PQS found in the human genome present high frequency of repetition -with just 72% being unique- and high chance of G4-formation, being a third of the total results over 40 in score. These values surpass most other species examined. However, some exceptions exist (Figure 4, B).

Figure 4. PQS-densities (PQS per 100000 nucleotides), % of unique PQS and number of known-to-form G4 (*in vitro*) sequences detected in humans (A.) and various other organism (B.). Text in red are the most outstanding results. White bars are the genomic PQS density found with method 2 (M2A), black bars are the genomic PQS density which score at least 40. Red dotted line is the M2A human (gh38) average PQS density, and green dotted line is the human (gh38) average PQS density which scores at least 40. Density is calculated by the formula: [(number of results of

Method)/(Total size of the genome)] x 100000. Non-unique % of PQS are calculated by the formula: [(number of results of M2A - number of results of M2B)/(number of results of M2A)] x 100, to give the percentage of sequences which have a frequency of occurrence bigger than 1.





On the one hand, *Leishmania* -and to a less extent the *Trypanosoma* and *Toxoplasma* genus- have very PQS-dense genomes with many known-to-form G4 sequences within. In *Leishmania major* for example, over 8000 PQS were detected containing the sequence 22Ag,(34) with the motif GGGTTA. Also more than 300 PQS containing T30695(35), and with less frequency T30177,(35) VEGF,(34) Scer21,(36) 26gsc,(37) Nef8528,(38) IV-1242540,(33)

CEB1,(39) CC,(40) C,(40) Bc,(40) B-raf,(41) A3T,(37) A,(40) 96del,(42) 27rap(37) and (TG5T)4,(34) were detected. In *T. gondii*, over 3000 sequences containing the G4-forming Ara24-1,(36) with the motif **GGG**TTTA, in addition to C,(40) Bc,(40) Chla27(36) and 93del(42) were also localized. On the other hand, *Plasmodium falciparum* and *Entamoeba histolytica* (causers of malaria and amoebiasis respectively) displayed very low PQS densities because of their high genomic AT content (80.6 and 75.2 % respectively). Still, the PQSs 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).

Gram-positive bacteria display very low genomic PQS 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 PQS densities in general for those studied here. *Pseudomonas aeruginosa* is the most outstanding genome in this group with a genome 3-fold denser than its human counterpart and with several known-to-form G4 sequences amongst the results. *Brucella melitensis* and *Neisseria meningitides* -causers of brucellosis and meningitis, respectively- follow next in density. Several confirmed known-to-form sequences were also found for *Treponema pallidum*,(43) indicating that G4s can also be interesting targets against syphilis.

The fungal genomes displayed lower PQS densities and uniqueness than the human DNA. *S. punctatis*, with a similar PQS density as the human's, is the densest of all examined and *S. cerevisiae* is the less unique with values also similar to the human. Many sequences known-to-form G4 *in vitro* were detected in all of these genomes, including Scer21*,(36) 26gsc*,(37) IX-356348,(33) IV-1242540(33) for *S. cerevisiae*, T30177(orI100-15)(35) and G4CT-pallidum(43) for *C. auris*, Tet22,(36) T30695*(35) and Nef8528*(38) for *C. albicans* and 22Ag**,(34) T30695,(35) Nef8624,(38) Cc,(40) Bc,(40) B for *S. punctatis*, where * is frequent: (freq. > 10) and ** is very frequent: (freq. > 100).

The viral genomes display a wide range of unique PQS densities. AIDS, hepatitis C, zika fever, rubella and dengue fever etiological causes have all higher densities than the human average. The most interesting result is the HIV virus which has an extremely high probable PQS sequences within its genome additionally to the known-to-form HIV G4 sequence PRO1.(44) Other viruses including the causers of ebola, flu, rabies and polio were totally void of PQS.

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 several scoring systems through machine learning together with its ability to locate already known-to-form G4 sequences makes G4-iM Grinder a practical and easy way of finding 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 looked for and analysed.

Using G4-iM Grinder, we examined the human genome to find new highly recurrent G-based structures as potential new G4 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. As example, we analysed HoEBR1 -a recurrent potential higher order quadruplex sequence with good score- and calculated the best combinations of PQS to form the structure.

A more macroscopic view of the human genome revealed chromosome 19 as the PQS densest chromosome and 13, 18 and Y as the least dense ones (with a fall of nearly 66 %). Our genome is still denser in PQS 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 G4 targets, as are the *Trypanosoma* and *Toxoplasma* parasites. In many of these organisms, we identified several sequences that have already been proven to form G4 *in vitro*. The bacteria and viruses inspected barely presented these known-to-form sequences because these differ from those confirmed in humans and listed in the G4Hunter and G4RNA databases. Still the pathological causers of AIDS, hepatitis C, rubella, zika and dengue fever show very high densities of unique PQS 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 or void of PQS and hence may require less stringent PQS search criteria to find potential targets (for example accepting G-runs of length 2). Surprisingly, some viruses have a high number of potential i-Motifs within their genome (data not shown) which may indicate an important role in their RNA and potentially G4 formation after transcription into the opposite strand. This is currently under study. Future work includes incorporating G4NN and Quadron as scoring systems and development of a shiny app for G4-iM Grinder.

DATA AVAILABITITY

The package and results of the human analysis can be found through GitHub ("EfresBR/G4iMGrinder"). Instructions on how to install the package can be located in the supplementary material section 6.

SUPPLEMENTARY DATA

Supplementary Data are available online.

ACKNOWLEDGEMENTS

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

FUNDING

This work was supported by The Spanish Ministerio de Economia y Competitividad (CTQ2015- 64275-P). Funding for open access charge: Spanish Ministerio de Economia y Competitividad.

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 tool. http://bioinformatics.ramapo.edu/QGRS2/index.php.

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

16. 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.

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. 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.

32. 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.

33. 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.

34. 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.

35. 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.

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

37. 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.

38. 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.

39. 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.

40. Dong,D.W., Pereira,F., Barrett,S.P., Kolesar,J.E., Cao,K., Damas,J., Yatsunyk,L.A., Johnson,F. and Kaufman,B.A. (2014) Association of G-quadruplex forming sequences with human mtDNA deletion breakpoints. *BMC Genomics*, **15**, 677.

41. 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.

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

43. 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.

44. 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: DNA and RNA G-Quadruplex, i-Motif and higher order structure search and analyzer tool

Efres Belmonte-Reche* and Juan Carlos Morales.

Content table:

1. Classification Performance of G4-iM Grinder (Method 2)

392 sequences from the supplementary material of G4Hunter were used to evaluate the classification performance of Method 2 of G4-iM Grinder using the recommended parameters. This list of sequences are 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 recognized as PQS. Only X3CGC with the sequence **GGG**CGC**GGG**CGC**GGG**CGC**GGG** was falsely recognized as a potential quadruplex with fairly high scores given by G4Hunter and PQSfinder (mean Score = 48).

The search on the G4-forming sequences showed that 233 of the 298 were correctly recognized by G4-iM Grinder when running predefined values. 65 were not recognized as PQS. There are several reasons why (Table 1):

1) Predefined search values require the presence of at least four G-runs to accept a structure as a PQS. Hence, G4s in the list that are intermolecular G4s (with less than 4 G-runs) did not get recognized when analyzed independently. However, most of these sequences were found in the human genome as part of bigger structures.

2) Predefined search values require G-runs of at least 3 guanines to be considered. Hence the structures that are composed of G-runs of size 2 did not get recognized.

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

4) Predefined search values require sequences to have loops no bigger than 10.

5) Predefined search values require sequences to have a total length of no more than 33. This can be eluded by also analyzing the sequences with Method 3 (size unrestricted search).

The predefined values can be easily modified if the user desires to adapt G4-iM Grinder to these G4s.

The analysis of the method 2 with the predefined variables, together with it's the 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 G4-Hunter and PQSfinder) of the 233 Sequences were then examined. The mean Score of all these structures was of 52.96. 75 % of them presented a score of 48 or more, and 97.9 % (228 of 233) scored more than 40. Hence, the filter used in this article to quantify the most probable to form PQS (Score > 40) score at least the same as 97.9 % of all verified to form sequences examined.

Table 1: Verified G4 sequences not recognized by G4-iM Grinder using Predefined search parameters. The reason why the Sequence was not recognized 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	GGGTGGGTTTTTTTTTTTTTTTGGGTGGG	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	GGGTGGGTTTTTTTTTTTTTTTTTTGGGTGGG	Loop > 10
Cppt2	TTTTAAAAGAAGGGGAGGAATAGGGGATATGA	Nº G-runs < 4
f3E3t	GGGTTTGGGTTTTTTTTTTTTTTTGGGTTTGGG	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	GGGTGGGTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTGGGTGGG	Loop > 10
X-106443	GGGTCCTCCAAGGGGTAAAACTTACATGGGATGGTGGGGTCACATGGG	Loop > 10 / Max PQS size > 33
f3S3t	GGGTTTGGGTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTGGGTTTGGG	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

2. 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
MinGruns: 4 *MaxPQSSize*: 20 *MinPQSSize*: 15	GGGTGGGTGGGTTTTAGGG GGGTTTTAGGGTTTTAGGG GGGTTTTAGGGTTTTAGGGTTTTAGGG	PQS Not-PQS Not-PQS
MinGruns: 2 *MaxPQSSize*: 10 *MinPQSSize*: 8	GGGTGGG GGGTTTAGGG GGGTGGGGGG	Not-PQS PQS PQS

3. Other variables:

Complementary: If TRUE, it will also calculate the complementary strand of the sequence and add it to the analysis. A new column will be created which will tell if the detected structures were found in the original (+) or complementary (-) strand.

DNA: As to tell the code if the Sequence is a DNA sequence. If *DNA* = FALSE, it will assume it is a RNA sequence. Useful for sequence complementary conversion and *KnownG4*.

Composition variables, both Method 2 and 3 can be configured to qualify the obtained final sequences.

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 %. 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 % of the final sequence (PQS/ i-motif) which is G, T, A, C and GGGTTA respectively.

KnownG4: for the found PQSs and only if *RunComposition* = "G" and if *DNA* = TRUE it will detect the presence of known DNA G4-forming sequences listed in the G4Hunter supplementary material.(1) If *DNA* = FALSE the PQS will be compared to those listed by Garant et al. in the DDBB G4RNA.(2) In both cases, these sequences were modified to start and end with the first and last G of the sequence to mimic the results structure of G4-iM Grinder. Also, resulting duplicated sequences were eliminated. In total, this resulted in 278 sequences of DNA and 232 sequences for RNA which have been demonstrated to form G4. As many names of the RNA sequences were repeated, a number corresponding to the row within the data table was pasted to the name, as to facilitate the identification of the G4. For example, WT/5 means it's the G4 called WT of row 5. This analysis will create a new column with the name/s of the detected sequence/s

followed by the nº of them detected in between brackets. This recount allows overlapping detection.

4. Scoring models and adaptations

G4-iM Grinder is capable of evaluating structures using 3 different algorithms published previously. cGcC and PQSfinder Scoring methods were adapted to fit into de searching methods developed in G4-iM Grinder whilst G4hunter was directly implemented from the code found in the supplementary material of the original article.

Even though all these codes were developed to find G-based PQS in DNA (G4Hunter and PQSfinder) or RNA (cGcC), it is really up to the user to use them in any other condition. G4Hunter and cGcC analyzes the relationship between G and C and hence it has potential of being of use in i-Motif evaluation. PQSfinder focuses on the tetrad size, bulges between tetrads and loop size to assess structures, which can also be potentially of use in C-based structures.

As G4Hunter punctuates negatively C presence, it was decided to make 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 bigger chances for the PQS to form real G4. On the other hand, bigger negative values will mean the sequence is more likely to form i-Motifs.

a. cGcC

Designed by Beaudoin et al.(3) to evaluate RNA PQS, it has also the potential of evaluating i-Motifs due to its algorithm being focused in G and C propensity. Such code follows the formula:

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

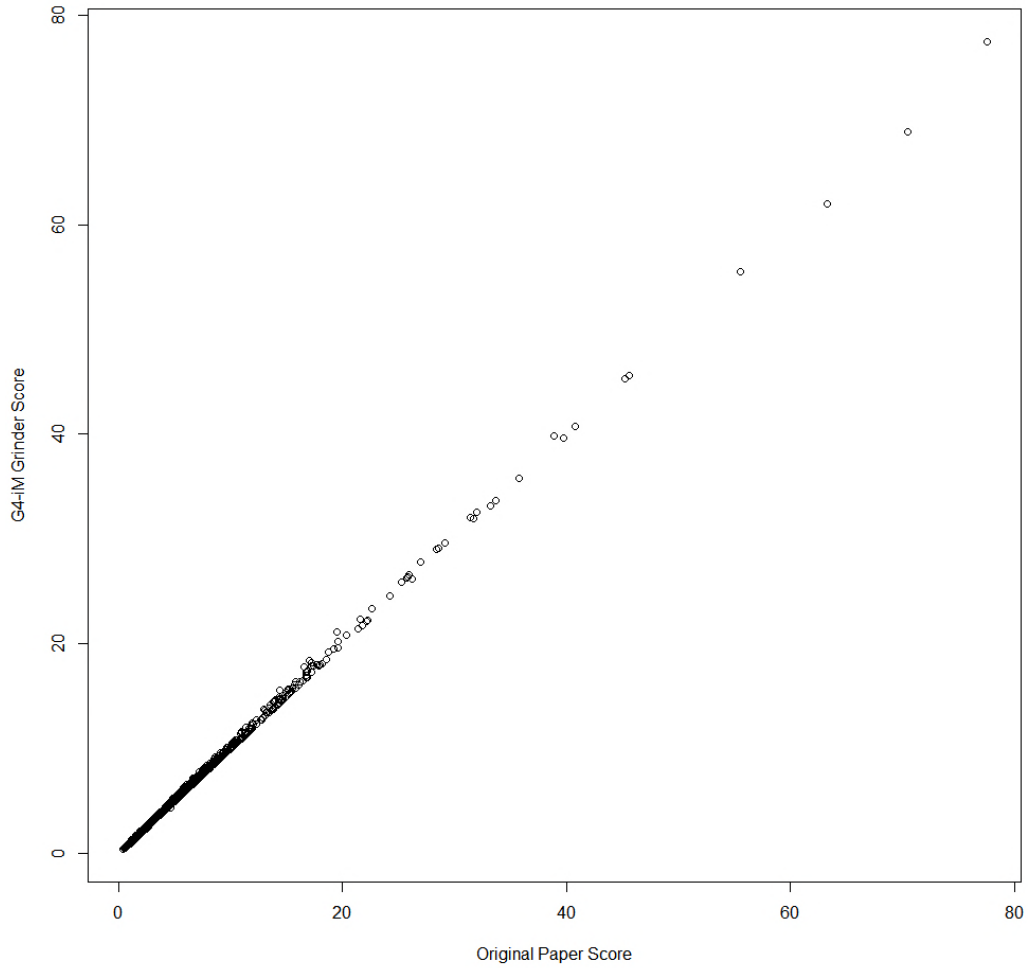
Where Gs(i) or Cs(i) is the number of G or C elements which compose a run, and i the number of runs. The analysis is done to a range of 50 nucleotides before and after the structures and the presence of complementary base structures diminishes the final score. An example is GGGCCCGGG which would score:

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

Given the nature of this scoring system, cGcC results are uncontained in the -100 to 100 range and hence we recommend using cGcC independently of other punctuation scores in the final global scoring value. For analysis of potential i-Motifs (*RunComposition* = C), the cGcC formula is elevated and multiplied by -1.

The >2000 sequences in the supplementary material found in the G4RNA article by Garant et al.(2) were compared with our results of 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(4) is based on several modulable constants which are applied to other factors, including: number of tetrads (Nt), number of bulges (Nb) and average loop size (Lm). Some modifications were made to the original formula as to adapt to the search parameters of G4-iM Grinder. Additionally, other changes were made to give greater freedom of analysis to the scoring system as the original functions allows only four G runs to be detected and evaluated, and ignores negative scoring sequences.

Hence to adapt the original formula, we simplified the appendix of the G run bulges penalization calculations involving *Fb* (bulge length penalization factor), *Lbi* (the length of the i-th bulge) and *Eb* (bulge length exponent) to 1. The variable number of tetrads (Nt) was substituted by the average run size (excluding from this calculation the bulge nucleotides) to better analyze longer sequences. The penalization for the bulges was modulated by adding the relationship between the minimal and the actual Nº of Runs in the Sequence. This was done to annul the dependency of this penalization factor towards the absolute number of bulges. As the other terms of the equation are based on averages (Loops) and fixed constants (tetrad size), punctuating larger PQSs with increasing numbers of bulges caused scoring distortions. Additionally, to all segments of the formula were a supplement constant added as to be able to modulate the response and approximate it to the original PQSfinder punctuation. For this reason, we also added exponential constants to the Tetrad and Loop segments. Hence the formula used was:

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

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

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

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

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

Given the changes on the formula, it was necessary to reevaluate the variable constants as to find those that give the nearest score to that of the original PQSfinder, and hence to their criteria of *in-silico* G4 formation. This was done by analyzing the non-chromosomal sequence of the human genome with G4-iM Grinder to detect all possible PQSs which fit Min and Max G runs size of 4 (as to allow original PQSfinder evaluation). When this was done, 7053 from the 18412 PQS found were discarded due to having no-score (negative scoring restrictions of the

original pqsfinder) and the remaining 11359 PQS were used to train the parameters. The results of this search are found in the table 2.

Table 2: PQSfinder variables. Both the original ones and the best fit conditions for G4-iM Grinder are shown, including the optimization tested range. Results of this process is shown last.

	Name	Description	Original Values	Recommended New Values	Range examined
Original Constants	*Bt*	Tetrad stacking bonus constant	40	14	10 to 50
	Pb	Bulge penalization constant	20	17	10 to 30
	Fm	Loop length penalization 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 upgraded system gave a mean difference between both scores for the same PQS of 4.92 (Original score = G4-iM Grinder score ± 4.92). 50 % of the 11359 PQS analyzed fell inside a ± 3 window, 75% of ± 6 and 95 % of ± 16 (Figure 2). Scoring variations arise in punctual PQS due to differences in definitions of G-runs, Loops and bulges and the differences from the formula.

However, similarities with the original system (R^2 = 0.89) are still more than enough to be able to predict effectively under the consensus of the original article those sequences with ample possibilities of forming stable structures (Figure 3).

When the run Composition is C, the algorithm will multiple the PQSfinder value by -1 as to make the value more negative, meaning bigger potential of forming *in-vitro* structures of i-Motifs.

Figure 2: PQS frequencies by their absolute differences between the original PQSfinder system and the implemented one in G4-iM Grinder for 11359 PQS. The horizontal red lines give the accumulative percentage of population which score inside the mentioned range. Graph done with ggplot2

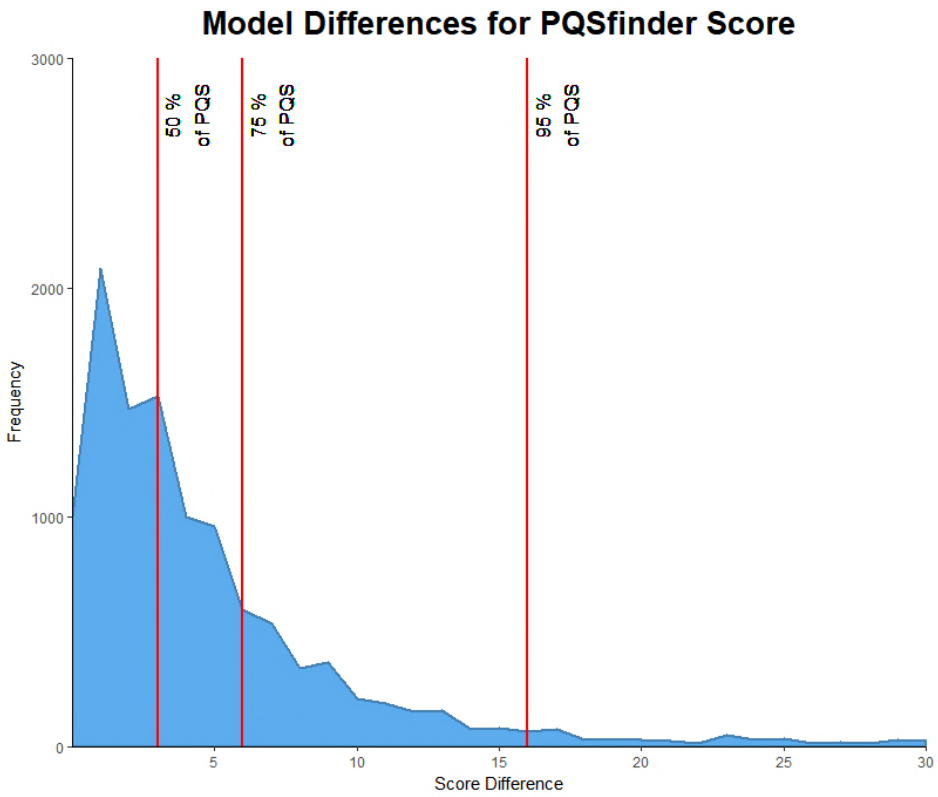
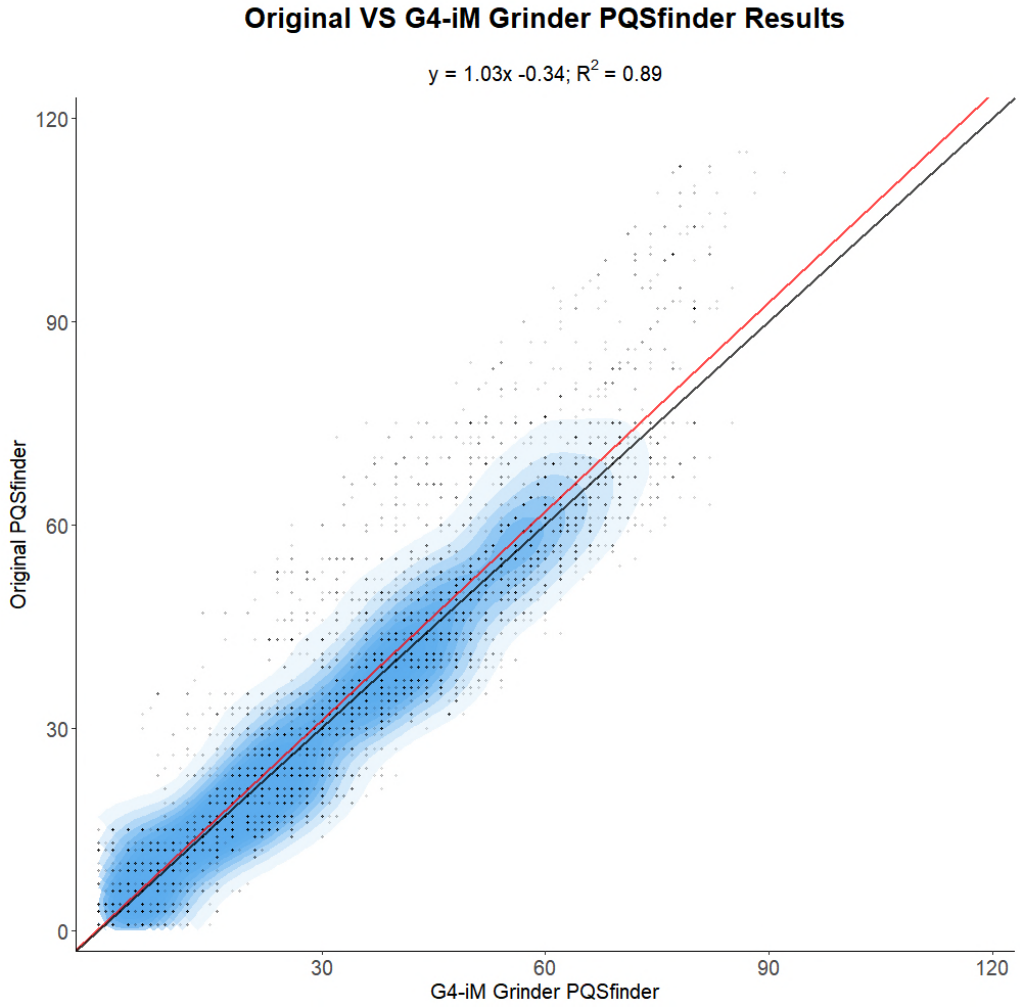


Figure 3: Relationship between the scoring of the original and G4-iM Grinder PQSfinder systems for 11359 PQS. The red line is the found correlation line (y = 1.03x - 0.34), and in black the expected perfect match one (y = 1x + 0). As a blue shadow, 2D PQS density to allow better contemplation of the relations between both systems regarding quantities. Graph done with ggplot2.



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 ranged from -4 to 4).

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

After applying all the desired scoring systems, a final score value which takes into account all these results can be calculated by means of a weighted average. To do so, the *WeightedParameteres* variable is used.

WeightParameters: is a vector of 3 numbers which gives the weight of each scoring system to calculate the final score of a PQS. The first value within the variables is given for the G4Hunter Score, the second to PQSfinder one 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* = c(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. It is recommended to interpret cGcC separately and not be included in weighted averages.

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

Score: For the results of method 2 and 3 and if at least 2 scoring systems are activated a final Score system will be calculated. To do so, a weighted mean value of the results will be given, where the weight of each value is defined within the variable *WeightParameters*. To calculate the score of the results for method 2b results, *FreqWeight* will be used to modulate the importance of the structures frequency.

For Method 2A and 3A:

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

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

Whilst for Method2B or Method3B, to this formula an appendix is added:

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

Where Freq. is the frequency of a PQS detected.

5. Second PHOQS analysis

The second example (Table 3) is a unique 125 nucleotide long sequence repeated 12 times every 300 nucleotides and exclusive to the end section of chromosome 10. This frequency is possibly higher as a big segment of unidentified nucleotides exists in between the continuous repetitions of the potential higher order PQS. These repetitions are part of the PPP2R2D gen which transcribes a regulatory protein phosphatase 2 subunit B (search done through the NCBI genome Data Viewer, December 2017). Dissection of the structure into its core PQS units (table 3 beneath entry 1 - extracted from M2A data) identified 16 possible overlapping PQS.

Table 3: Analysis examples of a higher-order PQS found exclusively in the + strand of human chromosome 10 (first row) and beneath are the possible PQS units that can potentially form the higher-order structure. In blue the G-runs, in red the loops, in green the bulges. G4Name are identified G4 sequences within the found structure which are known to form *in vitro*. Abbreviations: IL is total bulges, G4H is G4Hunter Score, PF is pqsfinder score, G, C, A and T is PQS composition which is that nucleotide (as %)

Freq.	Length	Runs	IL	PQS	G4H	PF	Score	G	C	A	T	G4Name
12	125	16	2	GGGAGGGAAGGCGGGCAGAGATGGAGAGAGGAACGGGAG ACCTAGAGGGGCGGAAGGATGGGCGGAGGGACGTTAGGA GGGAGGGAGGGAGGCAGGGAGGCAGGGAGGCAGGGAGG AACGGAGGG	35	44	**40**	59	10	27	4	
Possible PQS subunits												
I	26	4	1	GGGAGGGAAGGCGGGCAGAGATGGAG	35	36	**36**	62	8	27	4	
II	31	5	2	GGGAGGGAAGGCGGGCAGAGATGGAGAGAGG	33	28	**30**	61	7	29	3	
III	27	4	2	GGGAAGGCGGGCAGAGATGGAGAGAGG	30	18	**24**	59	7	30	4	
IV	33	5	2	GGGAAGGCGGGCAGAGATGGAGAGAGGAACGGG	30	26	**28**	58	9	30	3	
V	25	4	2	GGGCAGAGATGGAGAGAGGAACGGG	28	20	**24**	56	8	32	4	
VI	28	4	2	GGAGAGAGGAACGGGAGACCTAGAGGGG	29	22	**26**	54	11	32	4	
VII	26	4	0	GGGCGGAGGGACGTTAGGAGGGAGGG	41	53	**47**	65	8	19	8	
VIII	30	5	0	GGGCGGAGGGACGTTAGGAGGGAGGGAGGG	43	56	**50**	67	7	20	7	
IX	23	4	0	GGGACGTTAGGAGGGAGGGAGGG	43	56	**50**	65	4	22	9	
X	31	5	0	GGGACGTTAGGAGGGAGGGAGGGAGGG	42	55	**48**	65	7	23	7	
XI	19	4	0	GGGAGGGAGGGAGGG	51	60	**56**	74	5	21	0	
XII	27	5	0	GGGAGGGAGGGAGGCAGGGAGGCAGGG	47	58	**52**	70	7	22	0	
XIII	23	4	0	GGGAGGGAGGCAGGGAGGCAGGG	46	56	**51**	70	9	22	0	
XIV	31	5	0	GGGAGGGAGGCAGGGAGGCAGGGAGGCAGGG	44	55	**50**	68	10	23	0	
XV	27	4	0	GGGAGGCAGGGAGGCAGGGAGGCAGGG	42	52	**47**	67	11	22	0	
XVI	31	4	0	GGGAGGCAGGGAGGCAGGGAGGAACGGAGGG	40	48	**44**	65	10	26	0	

6. G4-iM 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, make sure 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 3.5.1 and R studio 1.1.456 or 1.1.463. 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")

#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)
## try http if https is not available
source("https://bioconductor.org/biocLite.R")
biocLite(c("BiocGenerics", "S4Vectors"), suppressAutoUpdate = TRUE,
suppressUpdates = TRUE)
```

To install and load G4-iM Grinder

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

Running some examples

```
# Using a genome available online
loc <- url("http://tritrypdb.org/common/downloads/release-36/Lmajor/fasta/TriT
rypDB-36_Lmajor_ESTs.fasta")
Name <- "LmajorESTs"
Sequence <- paste0(seqinr::read.fasta(file = loc, as.string = TRUE, legacy.mod
e = TRUE, seqonly = TRUE, strip.desc = TRUE, seqtype = "DNA" ), collapse = "")

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

# Executing a grind on the sequence in search of P. i-Motifs
Rs <- G4iMGrinder(Name = Name, Sequence = Sequence, RunComposition ="C")

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

7. 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, Dec 2013 – gh38) was download from the Sanger Institute (www.sanger.ac.uk) on July 2017.

The complete human G4iM-Grinder Results (737715 Kb) can be download from github: EfresBR/G4iMGrinder (https://github.com/EfresBR/G4iMGrinder) under "*Results human.RData*" and worked with in R. Result are organized in 25 lists and 1 data frame. 24 of these lists are the results for each individual human chromosome and the last (and biggest) is the complete and joint 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 comfortable way. More info on the meaning of these results can be found in the documentation of the package.

Other non-human genomes used in this paper were download on July/August 2018 from https://www.ncbi.nlm.nih.gov/, and include:

>	NC_001802.1	Human immunodeficiency virus 1
>	NC_012532.1	Zika virus isolate ZIKV/Monkey/Uganda/MR766/1947
>	NC_007373.1	Influenza A virus (A/New York/392/2004(H3N2))
>	NC_004102.1	Hepatitis C virus genotype 1
>	NC_001498.1	Measles virus,
>	NC_002549.1	Zaire ebolavirus
>	NC_003143.1	Yersinia pestis CO92
>	NC_002942.5	Legionella pneumophila subsp. pneumophila str. Philadelphia 1
>	NC_003197.2	Salmonella enterica subsp. enterica serovar Typhimurium str. LT2
>	NC_011750.1	Escherichia coli IAI39
>	NC_017548.1	Pseudomonas aeruginosa M18
>	NC_003997.3	Bacillus anthracis str. Ames
>	NC_003098.1	Streptococcus pneumoniae R6
>	NC_002737.2	Streptococcus pyogenes M1 GAS
>	NC_000962.3	Mycobacterium tuberculosis H37Rv
>	NC_002755.2	Mycobacterium tuberculosis CDC1551
>	NC_002677.1	Mycobacterium leprae TN
>	NC_004557.1	Clostridium tetani E88
>	NC_009495.1	Clostridium botulinum A str. ATCC 3502
>	NZ_CP003679.1	Treponema pallidum subsp. pallidum str. Sea 81-4
>	NC_011283.1	Klebsiella pneumoniae 342

>	NC_009386.2	Leishmania infantum JPCM5
>	NC_018228.1	Leishmania donovani BPK282A1
>	NC_001905.3	Leishmania major strain Friedlin
>	NC_004325.2	Plasmodium falciparum 3D7
>	NC_008409.1	Trypanosoma brucei brucei TREU927
>	NW_001848937.1	Trypanosoma cruzi CL Brener
>	NC_031467.1	Toxoplasma gondii ME49
>	NW_001916388.1	Entamoeba histolytica HM-1:IMSS
>	NC_032089.1	Candida albicans SC5314
>	NC_001133.9	Saccharomyces cerevisiae S288C
>	NW_015971537.1	Spizellomyces punctatus DAOM BR117BR117
>	NW_017263931.1	[Candida] auris strain 6684
>	NC_002505.1	Vibrio cholerae O1 biovar El Tor str. N16961
>	NC_001474.2	Dengue virus 2
>	NC_001611.1	Variola virus
>	NC_002058.3	Poliovirus
>	KP735609.1	Diachasmimorpha longicaudata rhabdovirus isolate UGA
>	NZ_LN831026.1	Corynebacterium diphtheriae genome assembly NCTC11397
>	NC_003317.1	Brucella melitensis bv. 1 str. 16M
>	NC_001545.2	Rubella virus
>	AL157959.1	Neisseria meningitidis serogroup A strain Z2491

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

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. 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.