1746–1759 *Nucleic Acids Research, 2016, Vol. 44, No. 4* *Published online 20 January 2016*
doi: 10.1093/nar/gkw006

Re-evaluation of G-quadruplex propensity with G4Hunter

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

[1]Université de Bordeaux, ARNA Laboratory, F-33000 Bordeaux, France, [2]Inserm U1212, CNRS UMR 5320, IECB, F-33600 Pessac, France and [3]CNRS-Université de Toulouse UMR5099, F-31000 Toulouse, France

Received October 7, 2015; Revised December 14, 2015; Accepted January 3, 2016

ABSTRACT

Critical evidence for the biological relevance of G-quadruplexes (G4) has recently been obtained in seminal studies performed in a variety of organisms. Four-stranded G-quadruplex DNA structures are promising drug targets as these non-canonical structures appear to be involved in a number of key biological processes. Given the growing interest for G4, accurate tools to predict G-quadruplex propensity of a given DNA or RNA sequence are needed. Several algorithms such as Quadparser predict quadruplex forming propensity. However, a number of studies have established that sequences that are not detected by these tools do form G4 structures (false negatives) and that other sequences predicted to form G4 structures do not (false positives). Here we report development and testing of a radically different algorithm, G4Hunter that takes into account G-*richness* and G-*skewness* of a given sequence and gives a quadruplex propensity score as output. To validate this model, we tested it on a large dataset of 392 published sequences and experimentally evaluated quadruplex forming potential of 209 sequences using a combination of biophysical methods to assess quadruplex formation *in vitro*. We experimentally validated the G4Hunter algorithm on a short complete genome, that of the human mitochondria (16.6 kb), because of its relatively high GC content and GC skewness as well as the biological relevance of these quadruplexes near instability hotspots. We then applied the algorithm to genomes of a number of species, including humans, allowing us to conclude that the number of sequences capable of forming stable quadruplexes (at least *in vitro*) in the human genome is significantly higher, by a factor of 2–10, than previously thought.

INTRODUCTION

Nucleic acids are clearly more than just a passive instruction manual for the cell. In addition to the protein-coding information content of nucleic acid sequence, the short-range secondary structures and long-range spatial organization, base modifications and chromatin accessibility are important components for cell functioning. Although most DNA in a cell is likely in the classical double-helix form (1) most RNA molecules have complex folds, reflecting their diversity of function. More and more light has been recently shed on alternative or unusual nucleic acids structures as sequences prone to such ‘unorthodox’ structures have been linked to a number of nucleic acid-related functions.

Guanine quadruplexes (G4) are a family of alternative nucleic acid structures that have attracted attention because of their high structural stability under physiological conditions and the widespread distribution of sequences compatible with G4 formation. The building block of the G4 structure is a guanine *quartet*, a planar association of four guanines stabilized by hydrogen bounds among the bases. The stacking of two or more quartets and the coordination of cations between them are responsible for the stability of the G4. Runs of G are a requirement for G4 formation by a given nucleic acid sequence. Identification of these G runs has been the foundation of tools to identify quadruplex forming sequences. The corpus of publications dealing with guanine quadruplexes has grown dramatically, and many of these publications provide *in vivo* evidence of quadruplex-related effects in telomere biology (2,3), transcription regulation (4), translation and RNA maturation (5,6), replication and genomic stability (7–9), and replication origin definition (10–13).

Several tools are available that predict quadruplex forming propensity. Seminal publications from the Balasubramanian and Neidle groups (14,15) describe the first generation algorithms that looked for patterns matching the stereotype [$G_nN_mG_nN_oG_nN_pG_n$] expected to be favourable for quadruplex formation. In a second-generation algorithm, the group of Maizels looked for the occurrence of runs of G_n ($n \geq 2$) in a window of a given size. Many vari-

[*]To whom correspondence should be addressed. Tel: +33 5 4000 3022; Email: jean-louis.mergny@inserm.fr
Correspondence may also be addressed to Laurent Lacroix. Tel: +33 5 6133 5992; Email: laurent.lacroix@inserm.fr
[†]These authors contributed equally to the paper as first authors.

ations have been proposed and applied to different types of genomic DNA or RNA sequence databases. These algorithms usually identify local enrichment of runs of G above a threshold size (*n*) of 2 or 3 (see (16) for review). Loop size has been (and is still) subject to discussion: initial studies constrained loop size between 1 and 7 nt, later developments allow longer loops, supported by experimental demonstration of formation of quadruplexes with loops of up to 30 nt (17).

A number of experimental studies established G4 formation for sequences that escape this consensus (for example (18)). These are *false negatives*. A more limited number of examples have been described of sequences that obey the consensus but do not form G-quadruplexes *in vitro* (17,19). These are *false positives*. Furthermore, many of these algorithms only provide a binary (yes/no; match/no match) answer rather than a quantitative analysis that would allow correlation of a given quadruplex 'strength' metric with other genomic or functional parameters.

To overcome these limitations, we chose to develop a new type of algorithm that takes into account G-*richness* and G-*skewness* of a given sequence and provides a score (quadruplex propensity) as an output. Richness reflects the fraction of Gs in the sequence and skew reflects G/C asymmetry between the complementary strands. We call this algorithm G4Hunter. To validate this model, we benchmarked it on a large dataset of 392 sequences from the literature (for example: (17,20)) or from unpublished results. We also validated this algorithm by analysis of the human mitochondria genome (16.6 kb) chosen because of its relatively high GC content and GC skewness as well as biological relevance of sequences with potential to form G4 near instability hotspots (21). The results of this search were validated using a combination of biophysical methods to accurately assess quadruplex formation of 209 sequences from the human mitochondrial genome *in vitro*. We then applied the algorithm to a number of species, including human, allowing us to conclude that the number of *bona fide* G4-prone sequences in the human genome must be very significantly re-evaluated. Our data suggest that the number of sequences in the human genome likely to adopt G-quadruplex structures is higher than previous estimates by a factor 2–10.

MATERIALS AND METHODS

Principle of the algorithm

In order to take into account G richness and G skewness, each position in a sequence is given a score between −4 and 4. The score is 0 for A and T (i.e., neutral or indifferent), positive for G and negative for C. To account for G-richness (or C-richness, meaning G-richness on the complementary strand), a single G is given a score of 1; in a GG sequence each G is given a score of 2; in a GGG sequence each G is given a score of 3; and in a sequence of 4 or more Gs each G is given a score of 4. The Cs are scored similarly but values are negative. This results in a near-zero average score for G-rich regions in GC alternating sequences that are likely to form stable duplexes that would compete with G4 formation. This scoring scheme also enables simultaneous scoring of the complementary strand. For a given sequence, the G4Hunter score (G4Hscore) is the arithmetic mean of this 'sequence' of numbers (Supplementary Figure S1A).

By construction, the G4Hscore is centred on 0 for random sequences, independently of GC content. This assumption was also verified on a number of genomes for which the sequence is not random. In contrast, the marked GC-*skewness* of the human mitochondrial genome leads to a non-null average score. The light C-rich strand (L strand) has a negative value of −0.4.

Genome-wide search

When analysing a genome-wide, the mean of the scored nucleic acid sequence is computed for a sliding window arbitrary set at 25 nt. Regions in which the absolute value of the mean score rises above a threshold are extracted. The overlapping region are then fused and refined by removing non-G (or non-C) bases at each extremity, which could have passed through the windowing threshold procedure. The sequence may also be extended if the first or last base is a G (or a C). To avoid cutting in a G (or C) run during the windowing threshold procedure, the previous and next bases, respectively, are also taken into account if necessary. The score of this fused and refined sequence is then computed again. This new score may be below the threshold when, for example, two fused sequences shared a G-rich core that compensates for two less favourable 'ends' (Supplementary Figure S1B). We decided to name G4H*x* the list of G4FS found by G4Hunter with a threshold of x (where *x* can in theory be between 0 and 4, but with typical values ranging between 1 and 2). G4-forming sequence (G4FS) density per kb is then calculated by dividing the total number of G4FS by the length of the genome. To take into account gaps in genomic sequence, bases that are given as N were not counted when evaluating genome size.

Scripting

G4Hunter scores were calculated either using Python or **R** scripts. Genome-wide searches were performed using an **R** script that takes advantage of existing packages such as GenomicRanges, rtracklayer, BSgenome, Biostrings and GenomicFeatures from BioConductor (22). Quadparser searches were also performed using a home-built version of a pattern searching script in **R** based on a previously published script (16). Receiver Operating Characteristic (ROC) analysis was performed using the ROCR package for **R** (23). Examples of script for **R** are provided as Supplementary Informations.

Genomes analysed

G4FS searches were performed on the human mitochondria genome (EF184640.1), 18 full genomes, including human, mouse, fruit fly and budding yeast, either from BSgenome packages for **R** (24) or from NCBI databases. All genomes analysed are listed in Supplementary Table S4. Script to generate bed or bigwig files allowing visualization in genome browser like UCSC genome browser or IGV genome browser are provided in the script 'ScriptG4Hunter.r' and detailed in the Supplementary Information.

Oligonucleotides

Oligonucleotides were purchased from Eurogentec and stored at −20°C. Oligonucleotide strand concentrations were determined using absorbance at 260 nm and extinction coefficients provided by the manufacturer. Sequences are provided in Supplementary Table S2.

Thioflavin I fluorescence assay

Thioflavin T (3,6-dimethyl-2-(4-dimethylaminophenyl) benzothiazolium cation) was purchased from Sigma-Aldrich (catalogue number T3516) and used without further purification. The fluorescence assay was performed as described previously (25).

Nuclear magnetic resonance

One-dimensional ^{1}H nuclear magnetic resonance (NMR) experiments were performed on a Bruker Avance 700 MHz spectrometer equipped with liquid TXI ^{1}H/^{13}C/^{15}N/^{2}H fixed-frequency probes and with Z- gradient probe. Analyses were performed as described previously (26).

Circular dichroism spectroscopy

Circular dichroism (CD) spectra were recorded on a Jasco J-815 equipped with a Peltier temperature control accessory as described previously (26).

Absorbance spectroscopy

All spectra were recorded on a Uvikon XL spectrophotometer in 10 mM lithium cacodylate buffer (pH 7.2) at 3 or 4 μM oligonucleotide strand (except when stated otherwise) supplemented with 100 mM KCl.

Thermal difference spectra (TDS) were obtained by taking the difference between the absorbance spectra of unfolded and folded oligonucleotides that were recorded at high (>90°C) and low (4°C) temperatures, respectively in buffer containing 100 mM KCl. TDS provide specific signatures of different DNA structural conformations, provided that the structure is not too heat stable (a number of G4 structures do not melt at high temperatures) (27).

Isothermal difference spectra (IDS) were obtained as described previously in (25) by taking the difference between the absorbance spectra from unfolded and folded oligonucleotides. These spectra were respectively recorded before and after potassium cation addition (100 mM KCl) at 20°C. IDS provide specific signatures of different DNA structural conformations.

UV melting experiments to determine melting temperatures (T_{m}) of G4 structures were performed as previously described (28,29). G4 unfolding is typically associated with a decrease in absorbance at 295 nm, giving an inverted transition at this wavelength.

RESULTS

Application of G4Hunter to a reference dataset

We performed a literature search for sequences for which G-quadruplex formation was experimentally confirmed or disproven. We added our unpublished data and obtained a dataset of 392 sequences for which G4 formation propensity was known. The G4Hscore was calculated for all the sequences (Supplementary Table S1). Quadparser analysis was performed in parallel using G-runs length setting of 2 or 3 and loop lengths of 7 or 12.

We plotted the distribution of the G4Hscore separately for '**G4**' sequences and '**not-G4**' sequences (Figure 1A). G4Hscore was significantly higher ($P = 9.5 \bullet 10^{-43}$) for **G4** sequences than **not-G4** sequences with average G4Hscore values of 1.64 ± 0.46 and 0.16 ± 0.66, respectively. The significance of the observed distribution differences was evaluated using the non-parametric Wilcoxon rank-sum/Mann–Whitney U-test (null hypothesis: distributions are not different). As shown in the histogram of the score distributions for this reference dataset (Figure 1B), a threshold of 1 (dotted green line) resulted in a good discrimination of **G4** versus **not-G4** sequences.

In order to evaluate the quality of our scoring system, we performed an ROC analysis. The ROC curve is characteristic of a good estimator with pronounced convexity toward true positive results. The area under the ROC curve (AUC) estimates the accuracy with which the algorithm discriminates between **G4** and **not-G4** sequences to more than 0.96 (an area of 0.5 represents a non-discriminating value whereas an AUC of 1 indicates a perfect prediction). To estimate the threshold 'quality', we evaluated the ROC results for five different thresholds between 1 and 2 and the position of the ROC analysis for three settings commonly used for Quadparser. Use of a threshold value of 1 in G4Hunter provided both highest sensitivity and the highest sensibility (Figure 1C). All threshold values performed better than Quadparser with G-runs of 2, whereas Quadparser with G-runs of 3 had a lower false positive rate than did G4Hunter. This analysis revealed the bias in our 'reference' library toward structures that fit the 'classical' definition of quadruplexes: sequences with at least 4 runs of at least 3 Gs separated by up to 7 bases (called QP37). With thresholds of 1 or 1.2, G4Hunter recovered more 'true' G4FS (G4 Forming Sequences) than did Quadparser (281 and 252 respectively for G4H1 and G4H1.2 versus 196 for QP37), although the false positive rate was higher for G4Hunter (10.6 and 6.4% for G4H1 and G4H1.2 versus 1.1% for QP37). The main advantage of G4Hunter is to accept as positive sequences that do not fit in the 'classical paradigm' of Quadparser and thus this algorithm reduce the miss rate from 34% in QP37 to 5.7 and 15.4% for G4H1 and G4H1.2 respectively.

We next evaluated the precision by determination of the fraction of sequences that formed a G4 detected experimentally among the sequences found by Quadparser or with a G4Hscore above a threshold (Figure 1D). For all the settings of Quadparser or for threshold above 1 for G4Hunter, the prediction is very robust with a precision above 95%. We may thus conclude that G4Hunter is as good as Quadparser to predict G4 formation on this dataset, keeping in mind that this library is heavily biased towards G4FS (76% of library sequences) and towards G4FS that fit the 'classical' definition (50% of sequences).

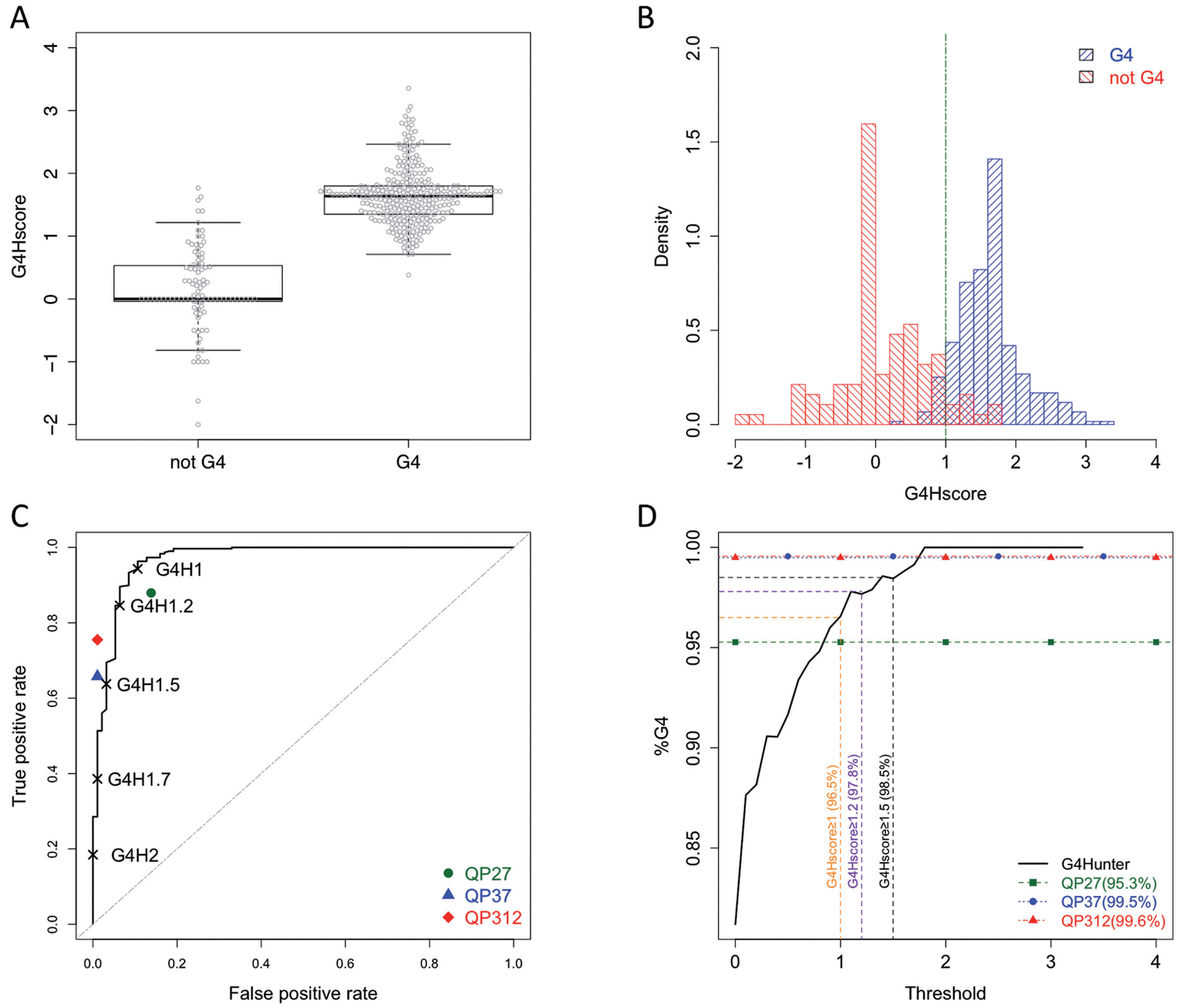

Figure 1. (**A**) Boxplot of the G4Hscore for the reference dataset. Opencircles represent the G4Hscore values for individual sequences belonging to either**G4** or **not-G4** classes. (**B**) Histogram of density distribution of the G4Hscores. Blue (right stripes) indicates **G4** and red (left stripes) **not-G4** forming classes, respectively. The dotted line indicates the value of G4Hscore for which more **G4** than **not-G4** sequences are found in this density histogram. (C) ROC curve for G4Hunter scores on the reference dataset. Black symbols represent the position of individual threshold values for G4Hunter. The symbols represent positions of the corresponding ROC values after applying Quadparser algorithm on the reference dataset with the following settings: runs of 2Gs and loop lengths between 1 and 7 (QP27, green dot), runs of 3Gs and loop lengths between 1 and 7 (QP37, blue triangle) and runs of 3Gs and loop lengths between 1 and 12 (QP312, red diamond). Random performing estimator would follow the dotted diagonal. (**D**) Precision versus threshold for G4Hunter. Fraction of sequence classified as G4 forming and which the G4Hscore is above the threshold in X-axis. Precision for the threshold 1, 1.2 and 1.5 are indicated with dotted vertical lines in orange, purple and black, respectively. Precision with QP27, QP37 and QP312 are indicated with the green squares, blue circles and red triangles lines respectively.

Application of G4Hunter to the human mitochondria genome

To evaluate G4Hunter on a less biased (or more accurately, differentially biased) dataset, we chose the human mitochondrial genome (release EF184640.1 from NCBI) because it is compact (16.6 kb circular DNA molecule) and GC-rich (and therefore potentially prone to G4 formation). The first eukaryotic genome to be sequenced (30), the mitochondrial genome encodes for mitochondrial subunits I, II, III, IV and V. The genome consists of a heavy (H) strand, rich in guanine bases and a complementary light (L) strand, which is rich in cytosines. The first G4 motif analysis was performed on the yeast mitochondrial DNA (31). Several studies found correlations between transcription termination, primer formation and R-loop stability and the presence of G4 structures (32). Computational analyses have correlated mitochondrial DNA deletions with non-B DNA propensity (33). Mitochondrial DNA deletions are observed in patients with human genetic disorders, and Quadparser-predicted mitochondrial G-quadruplex-forming sequences are located near known deletion breakpoints (21,34). A striking feature of the human mitochondrial genome is its GC skewness; this is reflected by the average G4Hscore value (−0.4).

We first calculated G4Hscores using sliding windows of 25 nt. We chose this window size as it is a close match to the mean length (26 nt) of the oligonucleotides in the G4FS reference dataset. Based on results from the analysis of the reference dataset, a threshold of 1 was chosen. Thus, we selected every sequence for which the G4Hscore was above 1 in absolute value in order to identify G4FS on both strand simultaneously. We identified 1846 overlapping sequences of 25 nt, corresponding to 170 windows of consecutive values that had scores above the threshold. As a direct consequence of the mitochondrial genome GC skewness, only four of these 170 regions (2.3%) are found on the L strand! From these sequences, 165 oligonucleotides were chosen for biophysical evaluation (Supplementary Table S2, see Supplementary Information for discussion of this choice).

In parallel, we searched the mitochondrial genome with Quadparser with the following settings: G-runs of 2 or more and loop lengths up to 7 nt (QP27), G-runs of 3 or more and loops up to 7 nt (QP37) and G-runs of 3 or more and loops up to 12 nt (QP312). These parameters identified 81, 5 and 11 candidates, respectively. All the hits found by Quadparser with G-runs of 3 and more were also found using G4Hunter. More than 50 sequences were found by G4Hunter and Quadparser with G-runs of 2 or more (Figure 2A). Twenty-six sequences were identified by Quadparser but not by G4Hunter; these sequences have G4Hscores below 1 and/or length <25 nt. The 26 sequences in this category are listed in Supplementary Table S2 as well as 11 sequences from the QP27 list which partially overlap motifs found with G4Hunter. For experimental validation, we also included seven sequences of 25 nt with G4Hscores between 0 and 0.4 that were expected to serve as negative controls.

We performed biophysical characterization of these 209 sequences. We initially assumed that a limited number of rapid assays would be sufficient for confirm G4 formation for a given sequence. However, demonstrating G4 formation *in vitro* was far more difficult than anticipated. In order to obtain a reliable answer to the seemingly simple question, '*Does this oligonucleotide form a quadruplex in vitro?*' we had to combine the results of up to six different techniques (IDS, TDS, UV-melting, CD, NMR and the thioflavin fluorescence assay). Taken individually, none of methods gave a satisfactory answer in all cases (IDS and NMR were the most reliable techniques). The experimental results for all sequences tested are presented in supplementary information ('Biophysical data' file): each page contains data on a single sequence.

Based on the conclusions from these tests, these sequences were classified as **G4** ($n = 71$), unstable G4 ('**UG4**'; $n = 63$) and **not-G4** ($n = 75$). The data were subjected to the same statistical analyses as were performed on the reference dataset. Results are summarized in Supplementary Figure S2A-F and discussed in the Supplementary Information. For these sequences, the G4Hunter discriminating score appears to be 1.2 (Supplementary Figure S2F). The G4Hscores are less dispersed than those from the reference dataset; this is not unexpected as most of the sequences tested were selected based on a score >1. This analysis indicated that for an actual genomic sequence, the G4Hunter score is a good estimator of the ability of a sequence to fold into a G-quadruplex when analysed *in vitro*. On the other hand, for sequences with a G4Hscore between 1.2 and 1.5, the conclusion is not so straightforward.

To evaluate the ability of G4Hunter to identify 'true' (at least *in vitro*) G4FS, we adjusted the G4Hunter results to avoid redundancy by merging overlapping sequences using thresholds ranging from 1 to 2. We obtained 96, 67, 25, 19 and 7 hits for threshold values of 1, 1.2, 1.5, 1.7 and 2, respectively. We then identified the overlaps between these sequences and the list of 209 sequences characterized biophysically. We used this set of results to generate the Euler diagram presented on Figure 2A taking into account sequences from the G4H1 list which overlap with more than one sequence from the QP27 list. This led us to reduce the QP27 list to 75 non-ambiguous regions. It appear clearly form this representation that sequences identified by both algorithm are more enriched in G4 than for each individual algorithm, but also the QP27 miss more G4 than G4H1 (13 versus 5). A comparison of the precisions of settings (Figure 2B) using also the intermediate class **UG4** (unstable G4) revealed that G4Hunter with a threshold of 1.2 results good precision with a low false discovery rate (FDR) of 10% compared to QP27 (FDR = 27%) if considering both stable and unstable G4. But if only high confidence (stable) G4 are seeked (excluding **UG4**, see Supplementary Information for a discussion regarding this point), the threshold for G4Hunter has to move up to 1.5 to get a FDR < 10%. This nevertheless comes to the cost of a reduced number of sequences identified. If an even lower FDR is required, increasing the threshold to 1.7 eliminate (for this dataset) all false positive result. On the other hand, QP312 and QP37 are also very good in ignoring false positive, but the results then miss half (for QP312) or three-fourth (for QP37) of the results found with G4H1.7.

We studied the impact of window size on G4Hunter performance. We computed the G4Hscores using windows ranging in length from 15 to 100 nt. The lower limit of 15 nt corresponds to the length of one of the shortest known stable intramolecular G-quadruplexes that is formed by the thrombin binding aptamer (TBA, d-GGTTGGTGTGGTGG). Most of the intramolecular quadruplexes are much shorter than 100 nt; the 100-nt upper value was chosen to match the window size of 100 nt chosen by Maizels *et al*. (35). In the mitochondrial genome, the highest number of hits was obtained with a window size of 15 and a threshold of 1, but these parameters resulted in an FDR of 68% (Supplementary Table S3). Increasing window size and threshold increased the precision of the detection, but reduced of the number of sequences identified. For a precision of 90% (i.e. FDR = 10%), the best results were found for a window size of 30 and threshold of 1.4 or for a window size of 20 and threshold of 1.7. If we require an FDR of 5%, the best settings were window size of 30 and threshold of 1.4 and with window size of 25 and threshold of 1.6. As most of the reported G4FS reported have a size around 25, we decide to use a window size of 25 nt with threshold values between 1 and 2 (a higher threshold means the sequences identified likely form G-quadruplexes, but at a cost of more false negatives) for genome-wide studies.

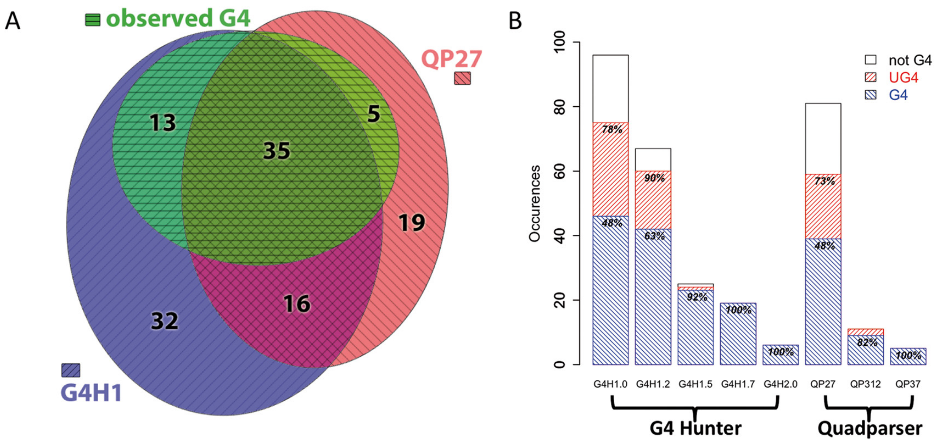

Figure 2. (**A**) Euler diagram representation of sequences from thehuman mitochondrial genome found by G4Hunter with a threshold of 1 (G4H1, blue, right stripes), sequences found by Quadparser (runs of 2Gs and loops length between 1 and 7, QP27, red, left stripes), and sequences experimentally demonstrated to form a G4 (green, horizontal stripes). Numbers indicate population of each subclass. (**B**) Number of sequences found by the different algorithms using various settings in the mitochondria genome (**G4**, in left striped blue, **not-G4**, in white and unstable G4 (**UG4**) in right striped red). The percentages in the blue bar indicate percentage for which G4 formation was experimentally confirmed. The % in the red bar indicate the fraction for which the conclusion of the biophysical test was **G4** or **UG4**. The number of sequences for each list in this panel is the number of non-overlapping sequences.

Analyses of genomes

Using G4Hunter, we identified G4FS in 20 different genomes including *Homo sapiens*, *Mus musculus*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Plasmodium falciparum*, *Escherichia coli* and *Arabidopsis thaliana*. We also included *Dictyostelium discoideum*, which has a GC poor (22%) genome. The number of potential G4FS is of course dependent on genome size (Supplementary Table S5A and B). To allow meaningful comparisons among species, we computed G4FS per kb. The occurrence of G4FS is comparable and the highest among mammals, ranging from 2.5 ± 0.1 per kb with a threshold of 1 (likely G4FS) to 0.17 ± 0.04 per kb with a threshold of 2 (very stable/highly likely G4FS) (Tables 1A and 1B). The relationship between G4FS density and threshold can be described with an exponential fit (Figure 3) with a similar exponential factor (slope) of −2.71 ± 0.25 for all mammalian genomes. Unsurprisingly, the human, chimpanzee and macaque genomes have similar densities and exponential factors for the dependency on threshold (−3.00 ± 0.07). Outside the mammalian class, the densities of G4FS and threshold dependencies are quite diverse. Four families of curves were observed (Figure 3): (i) G4FS-rich genomes (mammals, rice, chicken; Supplementary Figure S3A); (ii) intermediate G4FS genomes (*E. coli*, fly, bee and zebrafish; Supplementary Figure S3B); (iii) low G4FS genomes (budding and fission yeasts, *C. elegans* and *A. thaliana*; Supplementary Figure S3C); and (iv) very poor G4FS genomes (*P. falciparum* and *D. discoideum*; Supplementary Figure S3D).

For the first three classes, an exponential fit gives a very good description of the density versus threshold distribution. Interestingly, the slope varies considerably: from −2.4 for *A. mellifera* to −5.6 for *E. coli*. The GC-poor organisms *D. discoideum* and *P. falciparum* exhibit the lowest density of G4FS. For these two genomes an exponential fit failed to describe the density versus threshold curve with a marked deviation for thresholds above 1.4 (Supplementary Figure S3E). Above this limit, the G4FS density for *D. discoideum* decreased less markedly than did the density for *P. falciparum*. This could indicate a conservation of G4FS with high G4Hscores in *D. discoideum* arguing for a biological function in this organism. On the other hand, the marked decrease in G4FS density in the *P. falciparum* genome at high thresholds as well as the slopes for *E. coli*, both yeasts

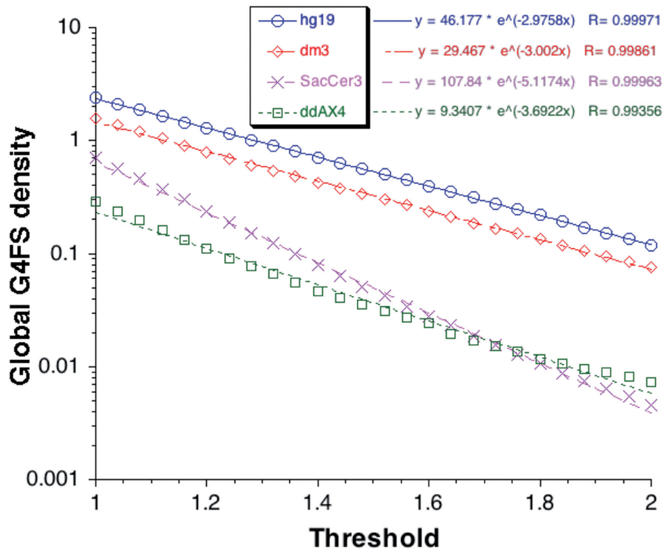

Figure 3. Global G4FS density versus threshold for four whole genomes. The number of hits found by G4Hunter using a window size of 25 was computed at different thresholds from 1 to 2 for *Homo sapiens* (hg19, blue circles), *Drosophila melanogaster* (dm3, red diamonds), *Saccharomyces cerevisiae* (SacCer3, pink crosses) and *Dictyostelium discoideum* (ddAX4, green squares) genomes. Data for other genomes are shown as Supplementary Information. The densities of hits per kb are represented with respect to the threshold used and was fitted using an exponential fit. The fitted equations are provided with the same code as the genome.

Table 1. Number of hits per kb of the sequenced genome obtained with G4Hunter using a window of 25 nt and the threshold indicated in the first column

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

Table 1A: Reference genomes are hg19 (Human), mm10 (Mouse), dm3 (*Drosophila*), ce10 (Nematode), ddAX4 (Dictyostelium), sacCer3 (*S. cerevisiae*), NCB.I20020305 (*S. pombe*), NCBI.20070724 (Plasmodium), Ecol.NCBI.20080805 (*E. coli*) and TAIR9 (Arabidopsis). Note that the *E. coli* reference genome contains 13 genomes of different *E. coli* strains.
In calculations of the lengths of the sequenced genomes, unattributed bases N were excluded.

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

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

and *A. thaliana* could indicate selection against very stable G-quadruplexes in these organisms.

The results of a G4Hunter analysis can be easily visualized with a genome browser like IGV (36) both at the promoter scale (Figure 4, top and Supplementary Figure S4A) or at a larger genomic scale (Figure 4, bottom and Supplementary Figure S4B). For the *MYC* promoter, we compared the G4FS identified using G4Hunter and Quadparser with different settings (Figure 4, top). Interestingly, some of the candidate sequences found with G4Hunter (with a threshold of 1.5 or more) were not found with the classical Quadparser parameters (QP37). At a larger scale (Figure 4, bottom), local enrichments in G4FS near *MYC* and *MIR1204* are striking. The profiles shown in Supplementary Information confirm that some genomic regions, such as the loci near *HRAS* and *SRC* (Supplementary Figure S4B), more G4FS are identified than in other regions of the genome. In the budding yeast genome, there are a paucity of G4FS (Supplementary Figure S4C), but we did identify a previously characterized G4FS (31). In the fruit fly genome, a local enrichment in G4FS near the heterochromatin region overlaps with a region recognized by a G4-specific antibody (37). There is also a local enrichment of G4FS in the rDNA cluster of the human genome as previously noted (38) (Supplementary Figure S4C).

Distribution of G4FS in the human genome

For the human genome, we performed a search using three window sizes (20, 25 and 30 nt), sizes that gave a good compromise between number of hits and precision for the mitochondria genome. In all three cases, the number of G4FS decreased exponentially with the threshold with similar slopes (Supplementary Figure S4A, hit number versus threshold). Based on the results of the analysis of the mitochondrial genome, a window size of 25 and a threshold of 1.5 results in precision above 90%, meaning that more that 90% of the sequences identified in the human genome with these settings should form G-quadruplexes. Window sizes and thresholds of 30/1.36, 25/1.52 and 20/1.72 resulted in similar numbers of hits (around 1.5 million). The majority of the hits are found by the three settings (respectively 75, 74 and 77%). Given that most G4-forming sequences described in the literature are longer than 20 nt, we used a window of 25. We analysed genomic features, such as promoters, coding regions, introns and exons, by computing three metrics: (i) the fraction of the feature containing one or more G4FS: (ii) the fraction of the G4FS found in a genomic feature and finally (iii) the local density of G4FS per kb in the genomic features.

In the hg19 release used, the entire genomic sequence is not present and there are large domains of unknown sequence. For example, the first 16 and 19 Mbs of chromosomes 22 and 14, respectively, have not been assigned. A similar 32-Mb gap is found on the Y chromosome. To evaluate enrichment, we thus adjusted the size of the genome to exclude all ambiguous bases (annotated as N). We also excluded the mitochondrial DNA from this analysis as its transcripts are not annotated in the UCSC Known Genes list. Thus, although there are 3.1•10^9 bases in the human genome, our analysis was performed on 2.86•10^9 bases. To evaluate metrics (ii) and (iii), we computed the total coverage of the feature on the genome by dividing the sum of the length of the feature by the size of the genome. This is referred to as the global background. To better reflect the actual genome composition, G4FS lists were randomized 1000 times in the available and annotated genome using an **R** script (gaps of more than 20 were avoided). This allowed us to generate 1000 lists with the same strand, width and chromosomal distribution as G4FS lists. We call this the resampled background.

Promoters. Using genomic annotation from UCSC Known Genes list, we defined a promoter region as the 1 kb upstream of a transcription start site (TSS) of an unambiguous transcript. We obtained 39 692 unique promoter regions.

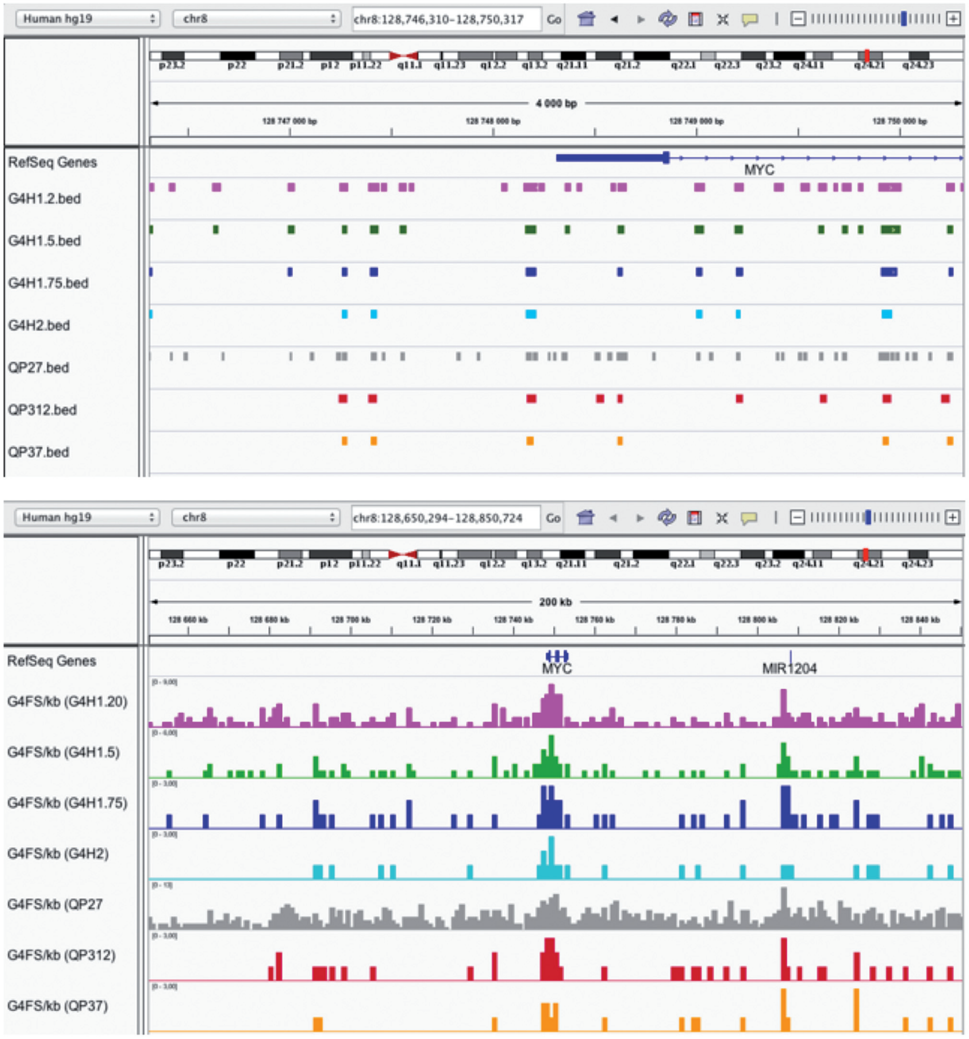

Figure 4. Genome browser views of the G4FS found near the *MYC* promoter. G4FS on a 4-kb region (top) and G4FS density on a 200-kb region (bottom) calculated by G4Hunter with thresholds of 1.2 (pink), 1.5 (green), 1.75 (dark blue) and 2 (light blue). G4FS from QP27, QP312 and QP37 are represented in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

(i) At a low G4Hunter threshold (1 or 1.2) or low stringency Quadparser setting (QP27), almost all promoters contain at least one G4FS (94, 84 and 94%, respectively). However, these percentages are close to those obtained for the mean of resampled backgrounds (93, 77 and 96%, respectively). Using more stringent criteria (G4Hunter with thresholds of 1.5, 1.75 and 2), 66, 52 and 37%, respectively, of promoters have at least one G4FS. These figures are significantly higher than the fractions obtained from the randomized lists (45, 27 and 14%, respectively, using these thresholds) and correspond to enrichments above the randomized background of 1.5-, 2- and 2.6-fold, respectively. Similar results were obtained with QP37 and QP312 (Table 2). Thus, the notion that promoters are enriched in G4FS is true when considering 'stable' quadruplexes only; these are the G4FS found by G4Hunter with a threshold above 1.5 or by Quadparser with blocs of at least 3 Gs (QP312 and QP37).

(ii and iii) When overlapping promoters are considered, promoter sequence represents 1.2% of the total human genome. If G4FS were random, one would expect that 1.2% of the G4FS would be found in promoters. This is not the case for the G4FS list extracted by G4Hunter with thresholds from 1 to 2 or by Quadparser: a 2.2–5.5-fold enrichment was found in all cases (Table 2). We also computed the occurrence of the hits in promoters for the resampled background for each list. The fraction of G4FS found in promoters by G4Hunter or Quadparser is 2–4.5-fold higher than the fraction of random sequences of the same chromosomal, length, and strand distributions with a p value $< 10^{-3}$ (p values were obtained from a null distribution based on resampled background). A similar conclusion was reached when the number of G4FS per kb (Table 2) were determined using either the global density of G4 in the genome (enrichment between 2.2 and 5.5) or the density computed by resampling (enrichment between 2 and 4.6).

This promoter enrichment, which is in agreement with previously published studies (35,39,40) also reflects the bias toward G/C or CpG island enrichment in promoters. Using the strand information and the position, we generated the

Table 2. G4FS in promoter region of the human genome

	G4H1	G4H1.2	G4H1.5	G4H1.75	G4H2	QP37	QP312	QP27
G4FS list size (no chrM)	*6 938 933*	*3 674 822*	*1436 253*	*707 090*	*339 975*	*361 982*	*706 788*	*8 572 750*
Promoter with G4FS ≥ 1 (%)	**94.1*****	**84.5*****	**66.5*****	**51.9*****	**36.9*****	**38.7*****	**52.7*****	**94.2**
Enrichment in promoter with G4FS ≥ 1 (resampled background, mean ± sd%)	*1.0 (92.8 ± 0.1)*	*1.1 (76.8 ± 0.2)*	*1.5 (45.4 ± 0.3)*	*2.0 (26.6 ± 0.2)*	*2.6 (14.0 ± 0.2)*	*2.6 (14.9 ± 0.2)*	*2.0 (26.8 ± 0.2)*	*1.0 (95.6 ± 0.1)*
Fraction GFS in Promoter (%)	**2.8*****	**3.5*****	**4.7*****	**5.6*****	**6.5*****	**6.8*****	**6.2*****	**2.8*****
Enrichment versus Global (global background,%)	*2.2 (1.2)*	*2.8 (1.2)*	*3.8 (1.2)*	*4.5 (1.2)*	*5.2 (1.2)*	*5.5 (1.2)*	*5.0 (1.2)*	*2.3 (1.2)*
Enrichment versus Resampled (resampled background, mean ± sd%)	*2.0 (1.4 ± 0.004)*	*2.5 (1.4 ± 0.006)*	*3.1 (1.5 ± 0.01)*	*3.7 (1.5 ± 0.014)*	*4.3 (1.5 ± 0.021)*	*4.5 (1.5 ± 0.021)*	*4.1 (1.5 ± 0.014)*	*2.0 (1.4 ± 0.004)*
G4FS/kb in Promoter	**5.38*****	**3.62*****	**1.88*****	**1.12*****	**0.62*****	**0.69*****	**1.23*****	**6.78*****
Enrichment versus Global (global background)	*2.2 (2.43)*	*2.8 (1.28)*	*3.8 (0.50)*	*4.5 (0.25)*	*5.2 (0.12)*	*5.5 (0.13)*	*5.0 (0.25)*	*2.3 (3.00)*
Enrichment versus Resampled (resampled background, mean ± sd)	*2.0 (2.71 ± 0.008)*	*2.5 (1.46 ± 0.006)*	*3.2 (0.59 ± 0.004)*	*3.8 (0.29 ± 0.003)*	*4.4 (0.14 ± 0.002)*	*4.6 (0.15 ± 0.002)*	*4.1 (0.3 ± 0.003)*	*2.0 (3.39 ± 0.01)*

The sizes of the different G4FS lists (G4Hunter with threshold of 1, 1.2, 1.5, 1.75 and 2, and Quarparser QP37, QP312 and QP312) were computed excluding the mitochondrial DNA. Percentages of promoters with at least one GFS were computed first by combining all the promoter regions of the Known Genes from UCSC (TSS-1000 to TSS) and then looking the presence of at least one G4FS in these regions. The enrichment was calculated using a resampled background. Fraction of G4FS located in promoters was computed by counting the occurrence of G4FS in a promoter region as defined before dividing by the size of the G4FS list. G4FS density (G4FS/kb) was computed by dividing the number of G4FS found in promoters with sum of the length of the promoter regions in kb. For these two metrics, enrichment was computed compared to a global background and to a resampled background. *** indicates $P < 10^{-3}$ for the null hypothesis that G4FS are randomly distributed on the chromosomes.

local profile of all G4FS within 1 kb of TSS (Figure 5 and Supplementary Figure S5C). A clear local enrichment between 200 bp upstream and the TSS was observed on both strands of the promoter. However, the most striking feature is the asymmetry between the coding and non-coding strands in the first 500 bp after the TSS with an enrichment on the coding strand. This could reflect an enrichment either in the post-TSS part of the promoter (i.e. in the genomic DNA) or in the 5′ UTR (i.e. in the RNA transcript). The enrichment in promoter regions may be a consequence of a transcriptional role for G4 structures or an indirect consequence of the typical properties of promoters such as the G/C richness, the presence of CpG islands and/or the presence of G/C-rich transcription factor consensus binding sites.

Other transcription-related genomic features. We performed a similar analysis for CDS, 5′ UTR, 3′ UTR, exons (unique, first, internal and last), introns (unique, first, internal and last), and in the 100-bp around intron/exon junctions. An obvious enrichment in G4FS was observed in the 5′ UTR, first exons and first exon/intron junctions (Supplementary Tables S6–S8) as previously observed (41–44). This confirms the observation from the distribution profile around the TSS and suggests a potential role of G4FS at the first splice junction in a pre-mRNA (41). The profiles of such splice junctions are shown in Figure 6 and Supplementary Figure S6. There is a local enrichment in G4FS on the coding strand in the intronic part of the splice junction.

Previous publication indicated that G4 potential correlates with gene function and that proto-oncogenes were enriched in G4FS while tumour suppressor gene were deprived in G4FS (35). Given the wide distribution of potential G4FS in promoters (from 37 to 94%), significance of gene function correlation with G4 is complex to analyse and would required a separate study. On the other hand, using a list of proto-oncogenes (n.ONC = 95) and tumour suppressor genes (n.TSG = 55) (35), we observed an enrichment in proto-oncogenes promoter for G4 found either G4Hunter with a threshold above 1.5 or with QP37 and QP312. Similar enrichments were also present for the first intron of proto-oncogenes for QP37, QP312 and sequences found with G4Hunter with a threshold above 1.75. The significance of these higher representation of potential G4FS was tested either with a binomial test compared to the ensemble of the UCSC Known Genes list or by choosing randomly 1000 times the same number of genes in the UCSC Known Genes list and using this distribution for the background G4 distribution. We failed to observe significant enrichment for the first exons, 5′ UTR or 3′ UTR of proto-oncogenes. Regarding tumour suppressor genes, significantly lower frequencies were observed for promoters and 3′ UTR when using a binomial test, but not when using a resampling approach (Supplementary Table S9). Further study and likely larger set of genes for each category would help to strengthen these conclusions but are out of the scope of this study.

DISCUSSION

As an increasing body of evidence suggests that G-quadruplex structures impact functions of nucleic acids in cells, there is a need for tools that accurately predict G-quadruplex forming propensity. Several algorithms are currently available, but a number of studies have established that there are regions of G4 formation that are not detected by these tools (false negatives) and that there are also a more limited number of examples of false positive detections (17,19). These observations prompted us to propose a new algorithm that relies on both G/C skewness and the presence of G-blocks. Rather than defining an arbitrary limit on the number of consecutive guanines required or on loop size, we chose to define a scoring function that reflects G-quadruplex propensity.

We experimentally validated the G4Hunter algorithm in an analysis of the human mitochondrial genome. This validation step turned out to be more complicated than expected: in order to reach reliable conclusions on the *in vitro* structures adopted by the hundreds of sequences investigated here, we had to rely on a set of up to six independent methods. Even then, these conclusions were somewhat ambiguous or subjective for 5% of the sequence tested. We suggest that a combination of techniques such as IDS, TDS, UV-melting, CD, NMR and a specific G4 ligand light-up fluorescence assay, should be used as the standard for experimental validation.

Whereas the Quadparser algorithm gives a ‘yes’ or ‘no’ answer, G4Hunter gives a score. The user may choose different threshold values in order to optimize the search. A high threshold (1.7 or more) will minimize the number of false

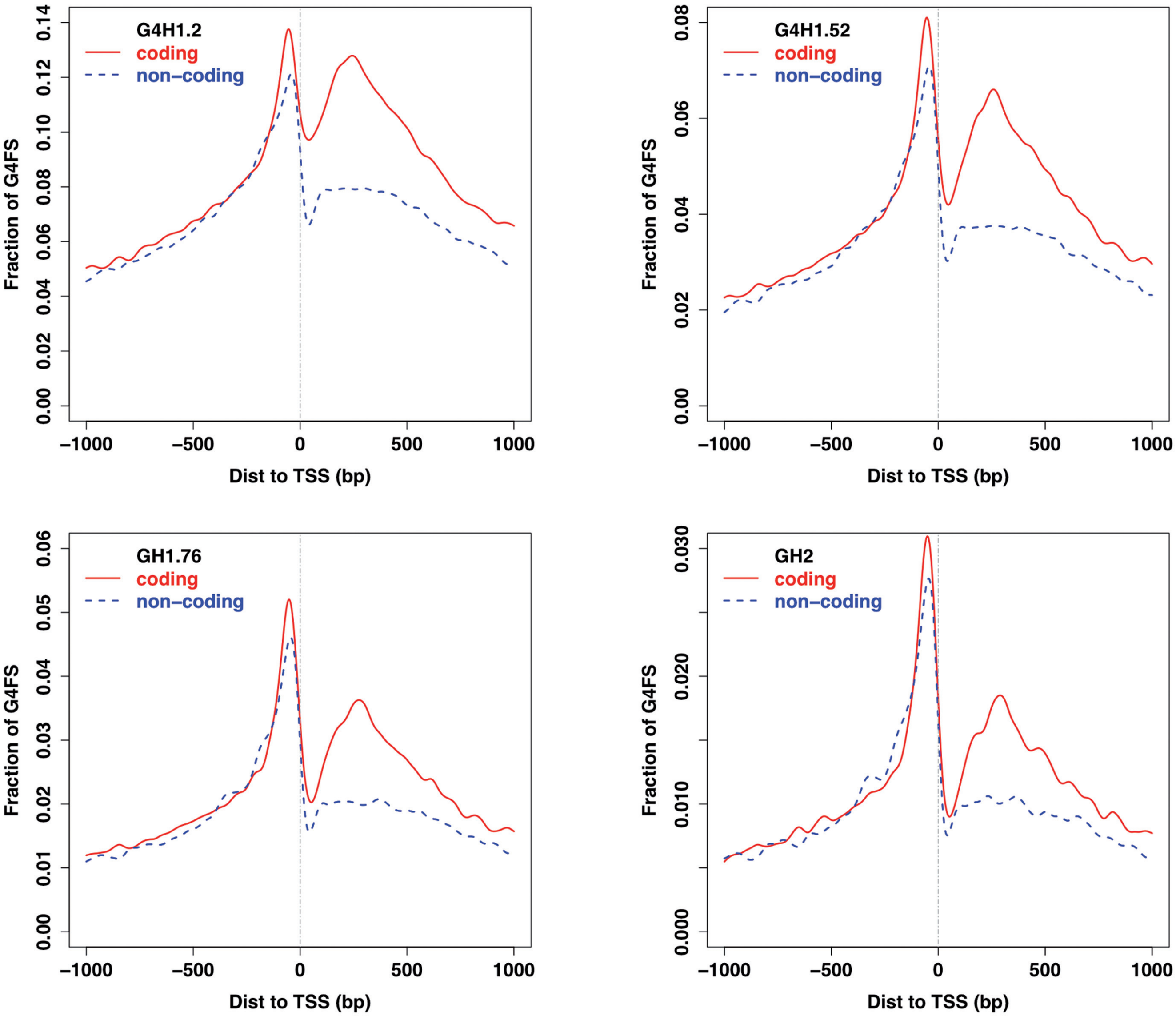

Figure 5. Profiles of G4FS around TSSs identified using G4Hunter with thresholdsof 1.2, 1.5, 1.75 and 2 (upper left, upper right, lower left and lower right, respectively). The Y axis is the fraction of G4FS at the nucleotide level. For each position the number of times this nucleotide is found in a G4FS was divided by the number of TSS sequences (39 692). The blue dotted and red solid curves correspond to the G4FS found on the non-coding and coding strands, respectively.

positive and favour highly stable G-quadruplex motifs but will miss a number of true G4-forming sequences. To obtain a more exhaustive survey of G4 potential, a lower threshold is recommended. A threshold of 1.2 is a good compromise, although some true G4 motifs will be missed. Choice of window and threshold will depend on the application: For example, in applications such as DNA origami or polymerase chain reaction (PCR) in which multiple oligonucleotide staples or primers are used, use of a threshold below 1.5 should minimize artefacts due to undesired G4 formation. On the other hand, when a structural analysis of a sequence is required (for example, after SELEX) a G4Hscore above 1.2 facilitate testing of the G4 hypothesis. A number of aptamers are known to adopt a G-quadruplex conformation, and our unpublished results (Kuznetsov *et al.*, in preparation) demonstrate that many of aptamers described in the literature form G-quadruplexes. Whatever the chosen threshold, one should remember that the discrimination is not perfect and that there will be both false positives and false negatives. As shown in Figure 1B, the discrimination issues are relatively limited.

Another important parameter is window size. Most of the analyses presented here were performed with the default window size of 25 nt. This corresponds to the actual size of many experimentally characterized G-quadruplexes. This parameter can easily be adjusted. Using a large window size (e.g., 100 nt) may actually aid in identifying genomic regions in which multiple contiguous G4 structures can be formed. This may well be relevant for some biological effects. For example, Rodriguez *et al.* found an increased

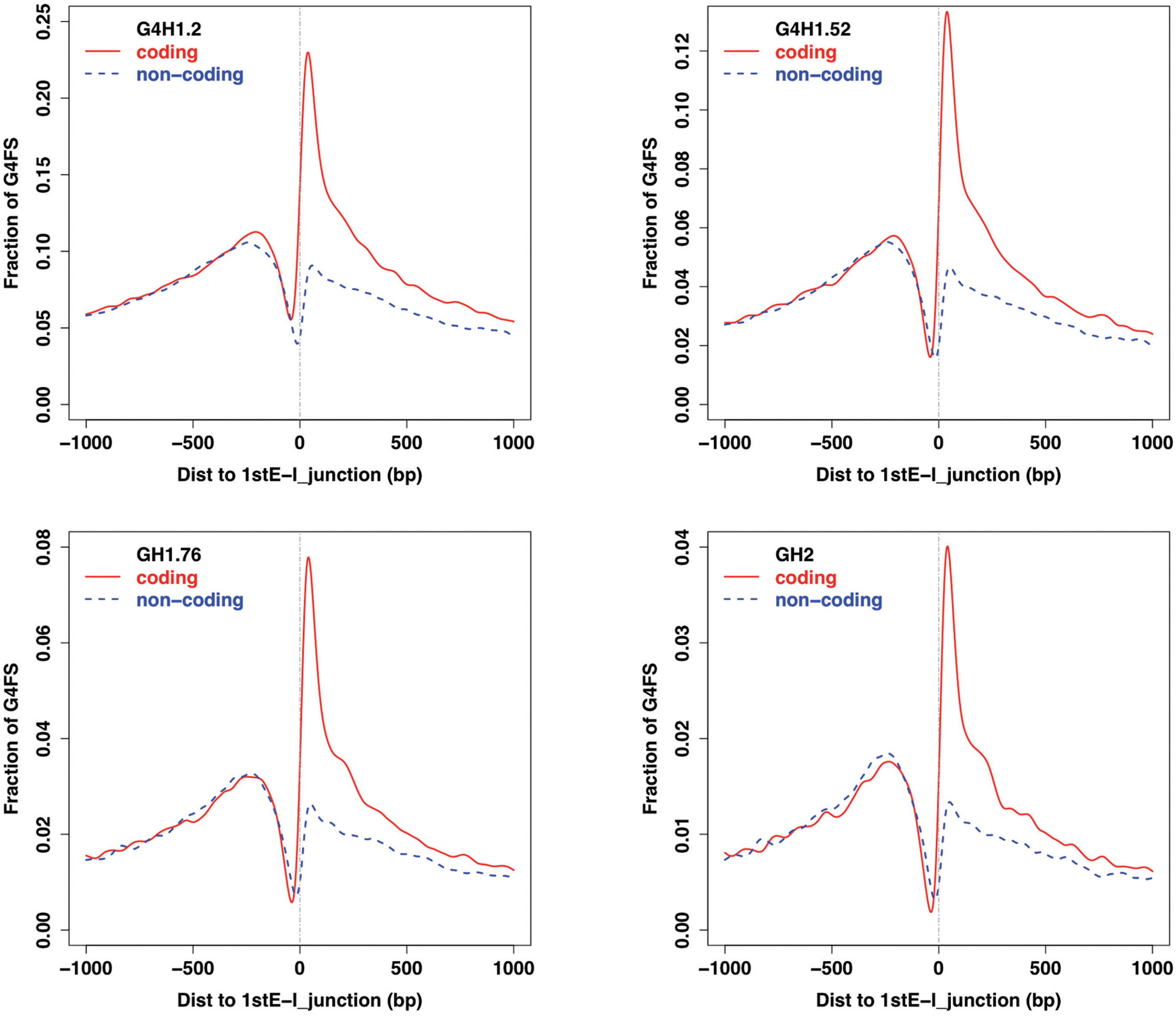

Figure 6. Profiles of G4FS around the first exon/intron junction for transcriptsin the UCSC Known Genes list with G4Hunter thresholds of 1.2, 1.5, 1.75 and 2 (upper left, upper right, lower left and lower right, respectively). The number on the Y-axis, the fraction of G4FS, represents at the nucleotide level for each position the number of time this nucleotide is found in a G4FS divided by the number of junction regions (37 466). The dotted blue and solid red curves correspond to the G4FS found on the non-coding and coding strands, respectively.

number of γH2AX foci in regions where multiple Quadparser hits are found (45).

This analysis raises a few specific points:

(i) *Double-stranded versus single-stranded nucleic acids*: the current version predicts whether an isolated strand forms or not a G4, assuming its complementary strand is transiently absent or sequestered. This is directly relevant for RNA sequences or for DNA primers, aptamers or origami staples, but needs to be placed in a genomic context for double-stranded DNA. The next step will therefore be to compare G4Hunter score with the calculated stability of the duplex formed with the complementary sequence. An interesting feature of G4Hunter is that its threshold may be adjusted. One may therefore chose conservative settings (a high threshold) to maximize the chances that the G4-predicted motifs are formed, even in a double-stranded genomic context.

(ii) *Molecularity*: in genome-wide studies, intramolecular structures are though to be more likely biologically relevant as these are the objects that DNA/RNA processing enzymes (polymerases, repair enzyme) will encounter in a monodimensional walk on their substrate. Nevertheless, DNA–RNA intermolecular hybrids may also be relevant (46) as well as complex G4 involving both strands of a DNA duplex (47). The search for such 'higher order' G4FS could be accommodated in G4Hunter by using a smaller window (<15) that would reduce the propensity to find intramolecular G4FS but still capture the the G-richness and G-skewness characteristic of such partial G4 motif, but such considerations are beyond the scope of this manuscript. In contrast, for *in vitro* applications such as PCR, DNA origami and SELEX, all molecularities should be considered. G4 formation may hamper applications such as PCR independently of whether the quadruplex is composed of 1, 2 or 4 strands. G4Hunter will mostly be used to identify intramolecular structures; however, the current search parameters and the experimental validation do not explicitly exclude species of higher molecularity. Search parameters can be slightly altered to favour intramolecular forms, for example, by imposing a minimum number of eight guanines within the window. With window sizes of 20 and 25 nt, a G7 tract embedded in a A/T rich region would give scores of 1.4 or 1.12, respectively, and therefore would be selected as a G4-prone motif, even though this stretch cannot form an intramolecular quadruplex. This feature can easily be implemented by adding a minimal number of G to be present for the sequence to be considered a hit. Thus, with a minimum number of eight Gs, this sequence would not be identified. This example also illustrates that short window sizes tend to make this problem more acute: the shorter the window, the more blatant the contribution of a single long G-run.

We compared the performance of G4Hunter with three different settings of Quadparser (QP37, QP312 and QP27). Interestingly, although stringent versions of Quadparser gave an excellent false positive rate (0 in the mitochondria dataset: i.e., all sequences predicted experimentally form G4 do), a large majority of true G-quadruplex forming motifs were missed. QP27 identified more true G4 but at the cost of accuracy (48%, Figure 2B). Interestingly, G4Hunter with a threshold of 1.2 has a higher precision (63%) with a comparable number of hits. In other words, the number of false positives is reduced with G4H1.2 as compared to QP27 (23 and 37 respectively).

Provided a reasonable G4Hscore threshold is chosen, the number of hits found at the genome wide level is higher than QP37. Given the data described here, the number of G4-prone sequences found in the human genome should therefore be significantly revised upwards compared to the commonly accepted figure of 376 000 G4FS. Remarkably, Balasubramanian *et al*. came up with a higher figure than this using the G4seq approach (44). To obtain a nearly identical number of sequences with G4Hunter, one has to select a threshold value of 1.75 with a window size of 25 (1.96 if the window size is reduced to 20). It is interesting to compare this figure with the number of sequences regions in the human genome that are predicted to be genomically unstable: (18,153) according to Nicolas *et al*. (48): a large majority of G-quadruplex-prone motifs are stable, arguing that such sequences may have regulatory functions without a concomitant deleterious instability.

Although G4Hunter finds far more likely G4 motifs than Quadparser with a lower false positive rate, our analysis demonstrated that G4Hunter has limitations. For example, values chosen for G_n blocks were integer values. This facilitates calculations when large genomes are analysed. However, there is no reason not to consider fractional values for these parameters, and we are currently investigating this possibility. Given our large database of sequences of known structure, we should be able to calculate accuracy for various values of *n*. Further, our scoring of A, T and C are context independent. This is obviously an approximation: for example, a GGGCGGGN…NGGGCGGG can form a very stable quadruplex with loops of single cytosine. The G4 formed is more stable with loops of single cytosine than with loops of single thymine or adenine (49). Our use of a negative score for C and null for A and T is not justified for this family of sequences. We therefore hope that our work will stimulate further studies aimed at improved G4Hunter!

G4Hunter was designed to take into account C with negative values both to disfavour regions rich in alternative G/C and to score both strand of a DNA duplex simultaneously. This may also allow G4Hunter to score i-motif forming sequences which we studied previously (50–52) and on which more and more groups are putting their focus (53–58). Rules dictating i-motif stability are still to be established but our previous works suggest that longer runs of C will be required (50), loops lengths will need to accommodate the presence of narrow and wide grooves. But this motif still required a C-richness and C-skewness that are the features on which G4Hunter is based. This could allow the characterisation of 'unusual' i-motifs with interrupted runs of C that are missed by motif-based search engines.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We are grateful to the Genotoul Bioinformatics Platform Toulouse Midi-Pyrenees for computing and storage resources. L.L. wishes to thank all the Cuvier team, in particular P.G.P. Martin, G. Micas, Y. Visser and W.T. Jones for support using R. P.G.P. Martin also participated in this work by providing code and advices regarding Bioconductor and R. J.L.M. and L.L. dedicate this work to the memory of Bruno Faucon and Ian Fraser Kilmister.

FUNDING

Fondation ARC (subvention libre to J.L.M.); Conseil régional d'Aquitaine Fellowship (to A.B.); *Agence Nationale de la Recherche* 'OligoSwitch' [ANR-12-IS07-0001], 'Quarpdiem' [ANR-12-BSV8-0008-01], 'VIBBnano' [ANR-10-NANO-04-03 to J.L.M.]; Université de Toulouse 'PISAAN' [AO1-2013 to L.L.]. Funding for open access charge: Inserm.

Conflict of interest statement. None declared.

REFERENCES

1. Watson,J.D. and Crick,F.H. (1953) Genetical implications of the structure of deoxyribonucleic acid. *Nature*, **171**, 964–967.
2. De Cian,A., Lacroix,L., Douarre,C., Temime-Smaali,N., Trentesaux,C., Riou,J.F. and Mergny,J.L. (2008) Targeting telomeres and telomerase. *Biochimie*, **90**, 131–155.
3. Zimmermann,M., Kibe,T., Kabir,S. and de Lange,T. (2014) TRF1 negotiates TTAGGG repeat-associated replication problems by recruiting the BLM helicase and the TPP1/POT1 repressor of ATR signaling. *Genes Dev.*, **28**, 2477–2491.
4. Siddiqui-Jain,A., Grand,C.L., Bearss,D.J. and Hurley,L.H. (2002) Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress c-MYC transcription. *PNAS*, **99**, 11593–11598.

5. Wieland,M. and Hartig,J.S. (2007) RNA quadruplex-based modulation of gene expression. *Chem. Biol.*, **14**, 757–763.
6. Millevoi,S., Moine,H. and Vagner,S. (2012) G-quadruplexes in RNA biology. *Wiley interdiscip. Rev. RNA*, **3**, 495–507.
7. Cheung,I., Schertzer,M., Rose,A. and Lansdorp,P.M. (2002) Disruption of dog-1 in Caenorhabditis elegans triggers deletions uptstream of guanine-rich DNA. *Nat. Genet.*, **31**, 405–409.
8. Lopes,J., Piazza,A., Bermejo,R., Kriegsman,B., Colosio,A., Teulade-Fichou,M.P., Foiani,M. and Nicolas,A. (2011) G-quadruplex-induced instability during leading-strand replication. *EMBO J.*, **30**, 4033–4046.
9. Paeschke,K., Capra,J.A. and Zakian,V.A. (2011) DNA replication through G-quadruplex motifs is promoted by the Saccharomyces cerevisiae Pif1 DNA helicase. *Cell*, **145**, 678–691.
10. Besnard,E., Babled,A., Lapasset,L., Milhavet,O., Parrinello,H., Dantec,C., Marin,J.M. and Lemaitre,J.M. (2012) Unraveling cell type-specific and reprogrammable human replication origin signatures associated with G-quadruplex consensus motifs. *Nat. Struct. Mol. Biol.*, **19**, 837–844.
11. Cayrou,C., Coulombe,P., Puy,A., Rialle,S., Kaplan,N., Segal,E. and Mechali,M. (2012) New insights into replication origin characteristics in metazoans. *Cell Cycle*, **11**, 658–667.
12. Valton,A.L., Hassan-Zadeh,V., Lema,I., Boggetto,N., Alberti,P., Saintome,C., Riou,J.F. and Prioleau,M.N. (2014) G4 motifs affect origin positioning and efficiency in two vertebrate replicators. *EMBO J.*, **33**, 732–746.
13. Comoglio,F., Schlumpf,T., Schmid,V., Rohs,R., Beisel,C. and Paro,R. (2015) High-resolution profiling of Drosophila replication start sites reveals a DNA shape and chromatin signature of metazoan origins. *Cell Rep.*, **11**, 821–834.
14. Huppert,J.L. and Balasubramanian,S. (2005) Prevalence of quadruplexes in the human genome. *Nucleic Acids Res.*, **33**, 2908–2916.
15. Todd,A.K., Johnston,M. and Neidle,S. (2005) Highly prevalent putative quadruplex sequence motifs in human DNA. *Nucleic Acids Res.*, **33**, 2901–2907.
16. Hershman,S.G., Chen,Q., Lee,J.Y., Kozak,M.L., Yue,P., Wang,L.S. and Johnson,F.B. (2008) Genomic distribution and functional analyses of potential G-quadruplex-forming sequences in Saccharomyces cerevisiae. *Nucleic Acids Res.*, **36**, 144–156.
17. Guedin,A., Gros,J., Alberti,P. and Mergny,J.L. (2010) How long is too long? Effects of loop size on G-quadruplex stability. *Nucleic Acids Res.*, **38**, 7858–7868.
18. Mukundan,V.T. and Phan,A.T. (2013) Bulges in G-quadruplexes: broadening the definition of G-quadruplex-forming sequences. *J. Am. Chem. Soc.*, **135**, 5017–5028.
19. Guedin,A., Alberti,P. and Mergny,J.L. (2009) Stability of intramolecular quadruplexes: sequence effects in the central loop. *Nucleic Acids Res.*, **37**, 5559–5567.
20. Stegle,O., Payet,L., Mergny,J.L., MacKay,D.J. and Leon,J.H. (2009) Predicting and understanding the stability of G-quadruplexes. *Bioinformatics*, **25**, i374–382.
21. Bharti,S.K., Sommers,J.A., Zhou,J., Kaplan,D.L., Spelbrink,J.N., Mergny,J.L. and Brosh,R.M. Jr (2014) DNA sequences proximal to human mitochondrial DNA deletion breakpoints prevalent in human disease form G-quadruplexes, a class of DNA structures inefficiently unwound by the mitochondrial replicative Twinkle helicase. *J. Biol. Chem.*, **289**, 29975–29993.
22. Huber,W., Carey,V.J., Gentleman,R., Anders,S., Carlson,M., Carvalho,B.S., Bravo,H.C., Davis,S., Gatto,L., Girke,T. *et al.* (2015) Orchestrating high-throughput genomic analysis with Bioconductor. *Nat. Methods*, **12**, 115–121.
23. Sing,T., Sander,O., Beerenwinkel,N. and Lengauer,T. (2005) ROCR: visualizing classifier performance in R. *Bioinformatics*, **21**, 3940–3941.
24. Pages, H. R package version 1.36.3 ed .
25. Renaud de la Faverie,A., Guedin,A., Bedrat,A., Yatsunyk,L.A. and Mergny,J.L. (2014) Thioflavin T as a fluorescence light-up probe for G4 formation. *Nucleic Acids Res.*, **42**, e65.
26. 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.
27. Mergny,J.L., Li,J., Lacroix,L., Amrane,S. and Chaires,J.B. (2005) Thermal difference spectra: a specific signature for nucleic acid structures. *Nucleic Acids Res.*, **33**, e138.
28. Mergny,J.L., Phan,A.T. and Lacroix,L. (1998) Following G-quartet formation by UV-spectroscopy. *FEBS Lett.*, **435**, 74–78.
29. Mergny,J.L. and Lacroix,L. (2009) UV Melting of G-Quadruplexes. *Curr. Protoc. Nucleic Acid Chem.*, Chapter 17, Unit 17 11.
30. Anderson,S., Bankier,A.T., Barrell,B.G., de Bruijn,M.H., Coulson,A.R., Drouin,J., Eperon,I.C., Nierlich,D.P., Roe,B.A., Sanger,F. *et al.* (1981) Sequence and organization of the human mitochondrial genome. *Nature*, **290**, 457–465.
31. 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.
32. Wanrooij,P.H., Uhler,J.P., Simonsson,T., Falkenberg,M. and Gustafsson,C.M. (2010) G-quadruplex structures in RNA stimulate mitochondrial transcription termination and primer formation. *PNAS*, **107**, 16072–16077.
33. Damas,J., Carneiro,J., Goncalves,J., Stewart,J.B., Samuels,D.C., Amorim,A. and Pereira,F. (2012) Mitochondrial DNA deletions are associated with non-B DNA conformations. *Nucleic Acids Res.*, **40**, 7606–7621.
34. 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.
35. 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.
36. Robinson,J.T., Thorvaldsdottir,H., Winckler,W., Guttman,M., Lander,E.S., Getz,G. and Mesirov,J.P. (2011) Integrative genomics viewer. *Nat. Biotechnol.*, **29**, 24–26.
37. Hoffmann,R.F., Moshkin,Y.M., Mouton,S., Grzeschik,N.A., Kalicharan,R.D., Kuipers,J., Wolters,A.H., Nishida,K., Romashchenko,A.V., Postberg,J. *et al.* (2016) Guanine quadruplex structures localize to heterochromatin. *Nucleic Acids Res.*, **44**, 152–163.
38. Drygin,D., Siddiqui-Jain,A., O'Brien,S., Schwaebe,M., Lin,A., Bliesath,J., Ho,C.B., Proffitt,C., Trent,K., Whitten,J.P. *et al.* (2009) Anticancer activity of CX-3543: a direct inhibitor of rRNA biogenesis. *Cancer Res.*, **69**, 7653–7661.
39. Huppert,J.L. and Balasubramanian,S. (2007) G-quadruplexes in promoters throughout the human genome. *Nucleic Acids Res.*, **35**, 406–413.
40. Du,Z., Zhao,Y. and Li,N. (2008) Genome-wide analysis reveals regulatory role of G4 DNA in gene transcription. *Genome Res.*, **18**, 233–241.
41. Eddy,J. and Maizels,N. (2008) Conserved elements with potential to form polymorphic G-quadruplex structures in the first intron of human genes. *Nucleic Acids Res.*, **36**, 1321–1333.
42. Huppert,J., Bugaut,A., Kumari,S. and Balasubramanian,S. (2008) G-quadruplexes: the beginning and end of UTRs. *Nucleic Acids Res.*, **36**, 6260–6268.
43. Eddy,J. and Maizels,N. (2009) Selection for the G4 DNA motif at the 5′ end of human genes. *Mol. Carcinog.*, **48**, 319–325.
44. Chambers,V.S., Marsico,G., Boutell,J.M., Di Antonio,M., Smith,G.P. and Balasubramanian,S. (2015) High-throughput sequencing of DNA G-quadruplex structures in the human genome. *Nat. Biotechnol.*, **33**, 877–881.
45. Rodriguez,R., Miller,K.M., Forment,J.V., Bradshaw,C.R., Nikan,M., Britton,S., Oelschlaegel,T., Xhemalce,B., Balasubramanian,S. and Jackson,S.P. (2012) Small-molecule-induced DNA damage identifies alternative DNA structures in human genes. *Nat. Chem. Biol.*, **8**, 301–310.
46. Wu,R.Y., Zheng,K.W., Zhang,J.Y., Hao,Y.H. and Tan,Z. (2015) Formation of DNA:RNA hybrid G-quadruplex in bacterial cells and its dominance over the intramolecular DNA G-quadruplex in mediating transcription termination. *Angew. Chem. Int. Ed. Engl.*, **54**, 2447–2451.
47. Cao,K., Ryvkin,P. and Johnson,F.B. (2012) Computational detection and analysis of sequences with duplex-derived interstrand G-quadruplex forming potential. *Methods*, **57**, 3–10.

48. Piazza,A., Adrian,M., Samazan,F., Heddi,B., Hamon,F., Serero,A., Lopes,J., Teulade-Fichou,M.P., Phan,A.T. and Nicolas,A. (2015) Short loop length and high thermal stability determine genomic instability induced by G-quadruplex-forming minisatellites. *EMBO J.*, **34**, 1718–1734.
49. Guédin,A., De Cian,A., Gros,J., Lacroix,L. and Mergny,J.L. (2008) Sequence effects in single-base loops for quadruplexes. *Biochimie*, **90**, 686–696.
50. Mergny,J.L., Lacroix,L., Han,X.G., Leroy,J.L. and Hélène,C. (1995) Intramolecular folding of pyrimidine oligodeoxynucleotides into an i-DNA motif. *J. Am. Chem. Soc.*, **117**, 8887–8898.
51. Mergny,J.L. and Lacroix,L. (1998) Kinetics and thermodynamics of i-DNA formation: phosphodiester versus modified oligodeoxynucleotides. *Nucleic Acids Res.*, **26**, 4797–4803.
52. Lacroix,L., Lienard,H., Labourier,E., Djavaheri-Mergny,M., Lacoste,J., Leffers,H., Tazi,J., Hélène,C. and Mergny,J.L. (2000) Identification of two human nuclear proteins that recognise the cytosine-rich strand of human telomeres in vitro. *Nucleic Acids Res.*, **28**, 1564–1575.
53. Day,H.A., Pavlou,P. and Waller,Z.A. (2014) i-Motif DNA: structure, stability and targeting with ligands. *Bioorg. Med. Chem.*, **22**, 4407–4418.
54. Brooks,T.A., Kendrick,S. and Hurley,L. (2010) Making sense of G-quadruplex and i-motif functions in oncogene promoters. *FEBS J.*, **277**, 3459–3469.
55. Lieblein,A.L., Furtig,B. and Schwalbe,H. (2013) Optimizing the kinetics and thermodynamics of DNA i-motif folding. *Chembiochem*, **14**, 1226–1230.
56. 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. (Camb)*, **48**, 10739–10741.
57. Khan,N., Avino,A., Tauler,R., Gonzalez,C., Eritja,R. and Gargallo,R. (2007) Solution equilibria of the i-motif-forming region upstream of the B-cell lymphoma-2 P1 promoter. *Biochimie*, **89**, 1562–1572.
58. Fujii,T. and Sugimoto,N. (2015) Loop nucleotides impact the stability of intrastrand i-motif structures at neutral pH. *Phys. Chem. Chem. Phys.*, **17**, 16719–16722.