Supplementary Information for
Reevaluation of quadruplex propensity with G4Hunter

Amina Bedrat, Laurent Lacroix & Jean-Louis Mergny

Discussion of the mitochondrial results:

From the 170 sequences selected with G4Hunter (window=25, threshold=1), 165 oligonucleotides were tested. This reduction arise from the fusion of some overlapping sequences. Out of the 81 sequences identified by Quadparser with the QP27 settings, 26 were non overlapping with sequences found with G4Hunter and 11 were partially overlapping. We decided to also test these 37 QP27 sequences.

The interpretation of the results from the combination of 5 to 6 methods was not always straightforward: for a few sequences, the conclusion whether G4 was formed or not was subjective and experimentator-dependent. We classified the sequences into three categories:

- **G4** (n=71), meaning that we are confident that G4 formation occurs for this oligonucleotide and this G4 is quite stable in KCl at 37°C;
- **unstable G4** (or **UG4**) (n=63), meaning that we have evidence for G4 formation but the G4 is unstable *per se* (mainly unfolded at 37°C) or in competition with another fold (based on UV profiles or NMR);
- **not G4** (n=75), meaning that we are confident that this oligonucleotide does not form a G4 under the condition tested.

Based on biophysical tests, we performed the same statistical analysis as the reference dataset (Supplementary Figure S2). The average G4H score was 1.36±0.33 for the **G4** class and 0.96±0.35 for the **not G4** and unstable G4 class (Supplementary Figure S2A). The difference was less pronounced than for the reference dataset but still significant according to a non-parametric Wilcoxon rank-sum/Mann-Whitney U-test (null hypothesis: distributions are not different), with a p of $3.1 \cdot 10^{-17}$. Using a less stringent definition for G4 formation and accepting unstable G4 as G4 results in reduced mean G4Hscores (**G4** (stable+unstable): 1.23±0.31; **not-G4**: 0.86±0.43) but the difference is still significative (Supplementary Figure S2B). Supplementary Figure S2C illustrate the distribution of the three classes if considered separately. From the histogram representation, the discriminating score appears to be 1.2 (Supplementary Figure S2D), but G4H scores are less dispersed than those in the reference dataset as most of the sequences were selected from the mitochondrial genome with a threshold of 1.

A ROC analysis (Supplementary Figure S2D) of the reference database illustrates that G4Hunter performs better than a random estimator on this dataset, but the accuracy reflected by the area under the curve (AUC=0.86) is less than that for the reference dataset, which by construction was biased for G4FS. From this analysis, G4Hunter appears to perform better the Quadparser for G4 structures with G-runs of 2 or more as the ROC curve is above this Quadparser point (green cross). The ROC analysis allowed us to propose that a threshold of 1.2 (triangle on Supplementary Figure S2D) is an excellent compromise between true positive rate (TPR) and false positive rate (FPR) as it is the point the farthest from the diagonal. This threshold results in a similar TPR as that obtained with Quadparser for runs of 2 Gs (0.66 *vs.* 0.63) but with a much reduced FPR (0.11 *vs.* 0.36). Using Quadparser with runs of Gs of 3 or more resulted in FPR of 0, comparable to G4Hunter with a threshold of 1.5 (FPR=0.007) or more but with a dramatically reduced TPR (0.08 and 0.14 for Quadparser QP37 and QP312 respectively *vs.* 0.27 for G4Hunter with a threshold of 1.5).

The ROC curve analysis allowed evaluation of the ability of the G4H score to determine which sequences form G4 in the whole population of potential G4 forming regions in the mitochondria genome but does not allow determination of the precision of the G4H score. This corresponds to the percentage of hits called by G4Hunter or Quadparser that form G4. This precision is 78%, 100%, and 100% for QP27, QP37, and QP312, respectively. For sequences with G4H scores above 1, the precision is 74%; it is 95% with a threshold of 1.2. With a threshold above 1.5, 100% of the hits form quadruplexes (Supplementary Figure S2E).

One has to be careful with these acronyms as even if FDR and FPR sound similar they both measure different properties: the FPR represents the fraction of false positive with respect to the whole negative population (Supplementary Figure S2G); the FDR represents the fraction of false positive with respect to the sequences identified as positive by the test (Supplementary Figure S2H). We also computed another metric called *accuracy* that measures the fraction of True Positive and True Negative in the population (Supplementary Figure S2I). This metric is greatly influence by the prevalence of the feature in the population (here 34% of the sequences tested form a G4) but is a straightforward estimation of the chance that the conclusion is correct. These three metrics illustrate that while QP37 and QP312 are seldom wrong when calling out a G4 and QP27 can quite often be wrong. G4Hunter for score between 1.1 and 1.6 is intermediate between QP37/QP312 and QP27 for the False Positives, but perfoms better for the TN and thus present a better accuracy.

Bioinformatic procedures:

The functions used for **R**-scripts are in the file "function_G4Hunter.r". The **R**-script "ScriptG4Hunter.r" provides examples of use of these functions. The file "ReferenceDataset.txt " contains the list of sequences used in the reference dataset.

The main script also contains instruction to generate a bed file that can be uploaded into a genome browser (IGV or the one from UCSC) to visualize the positions of the G4FS in a genomic context (see Supplementary Figure S4A and S4C for examples). It simply requires a GenomicRange produced by the G4hunt/G4huntrefined function and the function 'export' from the rtracklayer package in **R**. To generate the bigwig files for the G4FS density by kb, we used an home made function to calculate the number of G4FS per 1kb of non overlapping bin (G4GR2bigwig) and again he function 'export' from the rtracklayer package in **R**. Bigwig files are big and not easily upload into a remote UCSC genome browser but work perfectly with IGV genome browser (see Supplementary Figure S4B for an example). Bed files for all the data generate for this manuscript are also available upon request.

G4FS in genomic features:

Human genome annotation was performed using TxDb.Hsapiens.UCSC.hg19.knownGene package for **R** with modifications made to avoid ambiguity: redundant TSS have been fused, gene mapping at several loci discarded, and for introns and exons, we ensured that the first and last exons or introns were not found in other relative positions for other transcripts. The script used for these annotations is provided ("hg19annot.r"). Promoters were considered as the 1000 bp before the TSS. Intron/exon and exon/intron junction regions include 50 bp on each side of the junction. Examples of the scripts used to quantify presence of G4FS in a given genomic feature or fraction of feature with G4FS are provided in the script "GenoFeatScript.r". This script also illustrates the procedure used to plot the profiles around TSS as in Figure 5. The procedure used to generate the resampled background is also explained in this script. Typically, the genomic intervals corresponding to the G4FS were resampled keeping the chromosomal and strand distributions and avoiding large gaps of uncharacterized sequence, defined as sequences of only N for

more than 20 bases.

Supplementary Figure Legends

Figure S1:
A. Decomposition of the scoring procedure for a single sequence. The nucleotide sequence is first translated into numbers between -4 and +4 using the G4Hunter algorithm and the G4H score is calculated by taking the average of these numbers. Examples are shown using the following colour code: **G4** in red, i-motif in blue, **not-G4** (and not i-motif) in black. **B.** G4Hunter procedure to extract G4FS from long sequences. *(1)* The nucleotide sequence is first translated into numbers between -4 and +4 using the G4Hunter algorithm (G4Hunter_translate). *(2)* The G4H score is then computed in windows of size k (runmean in **R**, k=25). *(3)* Regions in which this runmean is above the threshold (here 1.5) are selected, and *(4)* the corresponding sequences are retrieved. The G4H scores of these sequences are calculated as in Supplementary Figue S1A. *(5)* Overlapping sequences are fused to avoid multiple counting of the same G4FS (overlapping sequences are indicated in blue) and the new G4H score is calculated. *(6)* Extremities of the sequences are processed to eliminate non G (or C) bases (in green) and also to retrieve Gs (or Cs) from the initial sequence that would have been removed by the windowing procedure (not shown in this example). Note that between step *(3)* and *(4)*, the extracted window sizes have been corrected using the runmean window size and thus the last two areas that were not overlapping at the runmean level are now overlapping at the nucleotide level.

Figure S2:
A. Boxplot of the absolute value of the G4H score for the mitochondrial dataset. Open circle represents the individual G4H score values for the two classes: **G4** and **not-G4** (unstable G4 are in this class). The *p* value for a Wilcoxon test between the two classes is indicated. **B.** Boxplot of the absolute value of the G4H score for the mitochondrial dataset. Open circles represent the individual G4H score values for the two classes: G4 (stable and unstable) and **not-G4**. The *p* value for a Wilcoxon test between the two classes is indicated. **C.** Boxplot of the absolute value of the G4H score for the mitochondrial dataset. Open circle represents the individual G4H score values for the three classes: **G4**, unstable G4 (**UG4**) and **not G4**. The *p* values for a Wilcoxon test between two classes are indicated. **D.** ROC curve for G4Hunter on the reference dataset. Black symbols represent the position of individual threshold values for G4Hunter. Green, blue, and red crosses represent 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), runs of 3Gs and loop lengths between 1 and 7 (QP37, blue), and runs of 3Gs and loop lengths between 1 and 12 (QP312, red). **E.** Precision *vs*. threshold for G4Hunter on the mitochondrial dataset. Fraction of sequences classified as G4 forming and for which absolute(G4H score) is above the threshold on the X-axis. Precision for the thresholds 1, 1.2, and 1.5 are indicated with dotted lines in purple, orange, and black, respectively. Precisions with QP27, QP37, and QP312 are indicated in green, blue, and red, respectively. **F.** Histogram of density distribution of the absolute values of the G4H score for the two classes of the mitochondrial dataset (blue, **G4** forming class; red, **not-G4** forming class). The green dotted line indicates the value of absolute(G4H score) for which more **G4** than **not-G4** are found in this density histogram. **G.** FPR *vs.* threshold plot. The lower part is a zoom for the threshold between 1 and 2. Purple vertical ligne indicates threshold of 1.2 (large dotted line) and 1.5 (small dotted line). **H.** FDR *vs.* threshold plot. The lower part is a zoom for the threshold between 1 and 2. Purple vertical ligne indicates threshold of 1.2 (large dotted line) and 1.5 (small dotted line). **I.** Accuracy

(ACC) *vs.* threshold plot. The lower part is a zoom for the threshold between 1 and 2. Purple vertical ligne indicates threshold of 1.2 (large dotted line) and 1.5 (small dotted line).

Figure S3:
Global density of hits per kb found by G4Hunter at thresholds between 1 and 2 and exponential fits for whole genomes. **A.** G4FS-rich genomes: hg19 (blue), rheMac3 (red), panTro3 (violet), mm10 (dark green), rn5 (green), canFam3 (brown), bosTau6 (yellow), susScr3 (orange), galGal4 (light blue), MSU7 (grey). **B.** Intermediate G4FS genome: *E. coli* (blue), fly (red), bee (violet), and zebrafish (green). **C.** Low G4FS genomes: budding (blue) and fission (red) yeasts, *C. elegans* (violet) and *A. thaliana* (green). **D**. Very poor G4FS genomes: *P. falciparum* (red) and *D. discoideum* (blue). **E.** Proposed decomposition for the fit of the very poor G4FS genomes (pink, *P. falciparum* and green, *D. discoideum*) for high and low thresholds compared to the human (blue) and budding yeast (red) fits.

Figure S4:
A. Genome browser view of the G4FS found by different algorithms near promoters. G4FS are represented for G4Hunter with the 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. Promoter regions of human *KIT*, *BCL2*, *KRAS*, *HRAS*, *SRC*, and *TERT* genes are on the top left, top right, centre left, centre right, bottom left, bottom right, respectively. **B.** Genome browser view of the G4FS density found by different algorithms in 200-kb region near the promoter represented in Supplementary Figure S4A. Results for G4Hunter with the thresholds of 1.2 (pink), 1.5 (green), 1.75 (dark blue), and 2 (light blue) and G4FS from QP27 (grey), QP312 (red), and QP37 (orange) are represented. *KIT*, *BCL2*, *KRAS*, *HRAS*, *SRC*, and *TERT* are on top left, top right, centre left, centre right, bottom left, bottom right, respectively. **C.** Genome browser view of the G4FS found on selected loci from budding yeast genome, fruit fly genome, and human rDNA cluster. G4FS are represented for G4Hunter with a threshold of 1.2 (pink), 1.5 (green), 1.75 (dark blue), and 2 (light blue) and for QP37 in orange. The budding yeast loci contains the sequences from Capra *et al.*, 2010 and the fruit fly loci are centred around the ANTP-C and BX-C loci from Hoffmann *et al.*, 2015.

Figure S5:
A. Number of hits found by G4Hunter for different thresholds between 1 and 2 for the human genome (hg19) with different window sizes (15, green; 20, purple; 25, red; 30, blue). The black line indicates the threshold that results in the same number of hits as with a threshold of 1.5 for a window of 25: 1.36 for a window of 30 and 1.72 for a window of 20. **B.** Overlaps between the hits found by G4Hunter on the human genome with a threshold of 1.52 and a window of 25 (k25_1.52), a threshold of 1.36 and a window of 30 (k30_1.36), and a threshold of 1.72 and a window of 20 (k20_1.72). The numbers within each area indicate the population of each subclass in thousands. **C.** Profiles of G4FS around the transcription start site (TSS) for the UCSC Known Genes list. G4FS lists used are, clockwise, G4H1 and three settings of Quadparser (QP27, QP312 and QP37). The Y-axis, the percentage of G4FS, represents at the nucleotide level for each position the number of times this nucleotide is found in a G4FS divided by the number of TSS region (39692). The blue and red curves correspond to the G4FS found on the non-coding and coding strands, respectively.

Figure S6:
Profiles of G4FS around the first exon/intron junction for transcripts in the UCSC Known Genes list. G4FS list used are, clockwise, G4H1 and Quadparser settings QP27, QP312, and QP37. The number on the Y axis, the percent of G4FS, represents at the nucleotide level for each position the number of times this nucleotide is found in a G4FS divided by the number

of junction regions (37466). The blue and red curves correspond to the G4FS found on the non-coding and coding strands, respectively.

Supplementary Tables:

Table S1: Reference dataset (excel file)

Table S2: Human mitochondrial genome dataset (excel file)

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

Table S4: Reference of the genomes used in this study

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

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

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

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

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

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

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

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

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

Table S9: Fraction of genomic features with at least one G4FS for a list of 95 proto-ongogenes and 55 tumour suppressor genes and statistical analysis compared to the whole USCS Known Gene List (genome) or to 1000 randomized list of 95 (cONC) or 55 (cTSG) genes (excel file).