SUPPLEMENTARY INFORMATION

MATERIALS AND METHODS

Non-B DB and STR annotations. Annotations of A-phased, direct, inverted and mirror repeats, G-quadruplexes and Z-DNA motifs were downloaded from the non-B DataBase (DB) at https://nonb-abcc.ncifcrf.gov. Additionally, we annotated STRs on the human reference (hg19) using STR-FM (*52*). We only considered mono-, di-, tri-, and tetranucleotide STRs with ≥8, ≥4, ≥3, and ≥3 repeats, respectively (*83*). We then collapsed STR motifs that could be matched by changing their reading frame (Table S9). For instance, $(AGC)_n$, $(CAG)_n$ and $(GCA)_n$ were collapsed into the $(AGC)_n$ group. We restricted our attention to non-B motifs and STRs annotated on autosomes.

Constructing genomic windows. Polymerization kinetics was studied in 100-bp windows (Fig. 1B). Motif-containing windows were centered at the middle coordinates of the annotated motifs in our list (Fig. S1). The centers of STRs with different repeat numbers were shifted to ensure their alignment (Table S9). Overlapping motif-containing windows (with motifs of the same or different type) were filtered out, leaving a total of 2,926,560 windows. All windows not containing motifs and not overlapping motif-containing windows were labeled as motif-free (a total of 3,649,152 windows).

IPDs. We used publicly available PacBio resequencing data (69x) from an individual male (HG002; NA24385) belonging to the Genome in a Bottle Ashkenazim trio (*53*). We analyzed 228 SMRT cells sequenced with P6-C4 chemistry in a mode maximizing the subread length and not the number of passes (analysis of the sequences obtained with P5-C3 led to similar results; data not shown) (*84*). On average, each molecule was sequenced in 2.12 passes, with the majority of the molecules sequenced only in a single pass resulting in a single subread (74.76% of the molecules). Sequencing reads were aligned to hg19 with pbalign (smrtanalysis-2.3.0), resulting in an ~52x average read depth, and IPDs were computed at

nucleotide resolution with *ipdSummary.py* (https://github.com/PacificBiosciences) — this produces one value per site averaging among at least three subreads, normalizing for inter-molecule variability and trimming for outliers. The resulting IPDs, which are strand-specific (any observed slowdown or acceleration of the polymerization concerns the strand used as template), were then used to populate motif-containing and motif-free 100-bp windows according to their coordinates (Fig. 1B); each window thus contains an IPD curve comprising 100 values or less (if some nucleotides lack IPDs). All windows with no IPD values were filtered out, and only motifs with ≥ 15 windows with IPDs on both strands were retained for subsequent analyses. This left us with a total of 2,916,328 motif-containing and 2,524,489 motif-free windows on the reference strand, and 2,916,377 motif-containing and 2,524,612 motif-free windows on the reverse complement strand. Next, for each motif type (Tables S2-S3), and separately for each strand, we aligned the 100-bp windows. This resulted in strand-specific IPD curve distributions for each motif type. An IPD curve distribution is visualized plotting quantiles (5^{th}, 25^{th}, 50^{th}, 75^{th} and 95^{th}) of the IPD values at each of the 100 nucleotides along the aligned windows (see Figs. 1B, 2A-D, S3, S5-S9). IPD distributions were visually unaffected by variants between the sequenced and the reference genomes.

Interval-Wise Testing for differences in IPDs. To detect statistically significant differences between IPD curve distributions in motif-containing and motif-free windows, separately for each motif type and strand, we employed the Interval-Wise Testing (IWT) procedure for "omics" data implemented in the R Bioconductor package and Galaxy tool *IWTomics (54, 55, 85)*. IWT treats the IPD values in a 100-bp window as a curve (see Fig. 1B) and assesses differences between two groups of curves (containing a given motif, and motif-free) performing a non-parametric (permutation) test at all possible scales, from the individual nucleotides to the whole 100-bp. When IWT detects a significant difference at a particular scale, it also identifies the locations (window coordinates) that lead to the rejection of the null

hypothesis (see Supplementary Methods for details). Because IWT is computationally expensive, we ran it on a maximum of 10,000 curves for each motif type and strand (sample sizes are listed in Tables S2-S3). For motif types with $n \geq 10{,}000$ windows, we randomly subsampled 10,000 windows and tested against a random set of 10,000 motif-free windows; this was repeated 10 times to ensure results robustness. For motif types with $n \leq 10{,}000$ windows, we tested both against a random set of 10,000 motif-free windows and against a random set of *n* motif-free windows; in both cases we repeated the comparison against 10 random sets, again to ensure results robustness. IWT was performed using three test statistics: the mean difference, the median difference, and the multi-quantile difference (i.e. the sum of the 5th, 25th, 50th, 75th and 95th quantile differences). Results for the latter, which most effectively captures differences in curve distributions, are presented in the main text (Fig. 2E), while those for mean and median are presented in Figs. S11C-D and S11E-F, respectively. P-values were computed using 10,000 random permutations (independent samples, two-tailed test). The procedure produces an adjusted p-value curve (comprising 100 p-values, one for each nucleotide, adjusted up to the selected scale) for each comparison (Fig. S2). We summarized results for all motif types in adjusted p-value heatmaps (Fig. 2E, multi-quantile difference; Fig. S11, mean and median). Red/blue indicate positive/negative observed differences, and are shown only for significant locations (adjusted p-value ≤ 0.05 in each of the 10 repetitions).

Effect of sequence composition on IPD. To investigate whether differences in IPD values depend on incorporation of different nucleotides, we computed mean IPD, a single nucleotide composition vector $P_{Si} = (p_A, p_T, p_C, p_G)$ $(p_A + p_T + p_C + p_G = 100\%)$, and a dinucleotide composition vector $P_{Di} = (p_{AA}, p_{AC}, p_{AG}, ..., p_{TT})$ $(p_{AA} + p_{AC} + p_{AG} + ... + p_{TT} = 100\%)$ in each 100-bp window. We considered only motif-free windows and combined data from both strands (results from the two strands considered separately were similar; data not shown). First, we measured the marginal effect of each nucleotide *j=A,C,G,T* as the

correlation between $log(mean\,IPD)$ and p_j. Next, we employed compositional regression models (*58*, *86*) to quantitate the overall effect of single nucleotide and dinucleotide composition on IPDs. The single nucleotide sequence composition vector P_{Si} was mapped to a three-dimensional euclidian vector $X_{Si} = (x_1, x_2, x_3)$ using the isometric log-ratio transform, and a multiple regression model was fitted for $log(mean\,IPD)$ on x_1, x_2, x_3. Model assumptions and validity were checked with standard multiple regression diagnostic plots and tests, and the R-squared was used to evaluate composition effect strength. Similarly, the dinucleotide composition vector P_{Di} was mapped to a 15-dimensional euclidian vector $X_{Di} = (x_1, x_2, ..., x_{15})$, and a multiple regression model was fitted for $log(mean\,IPD)$ on x_1, x_2, …, x_{15}. The dinucleotide compositional regression model fitted on motif-free windows, which had higher R-squared (see Results), was then used to predict the mean IPD values of motif-containing windows based on their composition, separately on each strand. For each motif type, we computed the differences between these predictions and observed mean IPDs, created their boxplots and performed two-sided t-tests for the mean difference being equal to zero — using a Bonferroni correction to adjust for multiple motif testing (Figs. 2F and S13).

Experimental characterization of G-quadruplexes. The ten most common G-quadruplex motifs (Table S5) from non-B DB annotations, as well as the $(GGT)_n$ motifs, were studied by circular dichroism (CD), native polyacrylamide gel electrophoresis (PAGE) and UV absorption melting profiles, as described previously (*87*). First, we considered only intramolecular G-quadruplexes, computed the mean IPD in each occurrence of the motifs, and fitted a simple regression for the mean IPD (on a log scale) on delta epsilon (for each motif delta epsilon was measured once, while mean IPD was computed for hundreds or thousands of occurrences; Table S5 and Figs. 3A and S15A). Next, we considered both intra- and intermolecular G-quadruplexes, and fitted a multiple regression for mean IPD (on

a log scale) on delta epsilon, the molecularity of the G-quadruplexes (either intra or intermolecular; a binary predictor), and their interaction. We fitted similar regressions replacing delta epsilon with melting temperature (T_m; Figs. 3B and S15B). In both cases we identified final models with backward selection.

SMRT sequencing errors. Data are again those from PacBio sequencing of HG002; NA24385 (*53*). Errors were analyzed restricting attention to motif occurrences (*not* motif-containing 100-bp windows). Due to potential misalignments at motifs in the repetitive parts of the genome, motifs and motif-free windows overlapping with RepeatMasker annotations (rmsk track obtained at https://genome.ucsc.edu) were excluded from this analysis. To focus on errors and not on fixed differences, all motifs and motif-free windows overlapping variants between HG002 and hg19 were also excluded (we used high confidence calls from a benchmarking dataset generated in (*53*)). For each motif type, control sets were constructed picking a filtered motif-free 100-bp window at random from within 0.5 Mb upstream or downstream of each motif occurrence, and trimming it to produce a motif-free region of the same length of the motif occurrence itself. This matches motif occurrences and motif-free regions in number and length (which guarantees the same measurement resolution for errors), as well as in broad genomic location (which accounts for megabase-scale variation in mutation rates across the genome (*64*, *70*). While not immediately relevant for error analyses, the latter is of importance for divergence, diversity and cancer somatic mutation analyses (see below), and thus was included here for consistency. We note that all results (for errors, divergence, diversity and cancer somatic mutations) are virtually unchanged if we do *not* match broad genomic location and select controls completely at random from the genome.

Error rates (the number of mismatches, insertions or deletions relative to hg19, divided by the total number of nucleotides from all subreads in a given region and expressed as a

percentage) were calculated for the newly synthesized strand that used six non-STR motif types and corresponding motif-free regions as a template. Since our purpose was detecting polymerase errors, we calculated the error rates based on individual subreads by accessing the alignment files directly and considering also low-frequency variants, including those supported by a single subread.

Illumina sequencing errors. We analyzed reads from the same individual male (HG002; NA24385) who was also sequenced with Illumina technology (*53*) using the previously subsampled data set with sequencing depth of 60x (*53*). Motifs and motif-free regions overlapping with RepeatMasker annotations or with variants between HG002 and hg19 were again excluded from the analysis. For reads aligning to the strand registered as reference sequence (hg19), we used a Naive Variant Caller (*88*) to detect variants supported by at least one read. These were separated into mismatches, insertions and deletions (compared to hg19). The analysis presented in the main text was limited to the newly synthesized strand of Read 1 that used our sets of motifs and motif-free regions as templates (results for Read 2, the other strand, and overall errors are provided in Supplementary Note 3). For any given motif occurrence and matching motif-free region (equal length, location within 0.5 Mb), read depths were extracted at each coordinate position and then summed across coordinates to obtain the total number of read bases sequenced for the region. Finally, error rates were computed dividing the number of variant calls by the total number of bases sequenced in each region.

Variants from human-orangutan divergence. We downloaded the 46 species Vertebrate Multiz Alignment (*89*, *90*) from the UCSC Genome Browser (Multiple Alignment Format (MAF) files from https://genome.ucsc.edu/index.html) and considered nucleotide substitutions between human and orangutan, as well as insertions and deletions in the human lineage after its divergence from the human-orangutan common ancestor (macaque

was used for the polarization of indels). These variants were intersected with our motif occurrences and matching motif-free regions (equal length, location within 0.5 Mb). For each motif, control was chosen at random from no further than 0.5 Mb upstream or downstream from the motif, to account for the effects of megabase-scale variation in mutation rates in the genome (*64*, *70*). To obtain an approximate measure of divergence, we divided the number of variants in each motif occurrence (or matching motif-free region) by their length. This was done separately for nucleotide substitutions, insertions and deletions. Motifs and motif-free regions overlapping with RepeatMasker annotations were excluded also from this analysis.

Variants from the 1000 Genomes project. We acquired all annotated variants from the 1000 Genomes project (Variant Call Format (VCF) files from http://www.internationalgenome.org/) and intersected the coordinates of those with a global frequency (across all populations) >5% with our motif occurrences and matching motif-free regions (equal length, location within 0.5 Mb). Indels were polarized using primate genomes (panTro4, gorGor3, ponAbe2 and nomLeu3) as previously described (*91*). To obtain an approximate measure of diversity, we divided the number of variants in each motif (or matching motif-free region) by their length. This was done separately for SNPs, insertions and deletions. Motifs and motif-free regions overlapping with RepeatMasker annotations were excluded also from this analysis.

Somatic mutations from The Cancer Genome Atlas. We acquired Annotated Somatic Mutations from the Genomic Data Common (GDC) Portal (VCF files from https://portal.gdc.cancer.gov/) and considered mutations identified by the MuTect2 software (*92*) for all tumor types available. MuTect2 uses sequencing reads from tumor and matched normal (cancer-free) samples to detect somatic variants with high confidence. It also discards variants that are likely to be sequencing errors (*92*). We used only the variants from the MuTect2 annotation passing all the softwares filters (FILTER ID=PASS). Variants in the

same positions from different files were considered as one (uniq). The resulting high-confidence variants were intersected with our motif occurrences and matching motif-free regions (equal length, location within 0.5 Mb). To obtain an approximate measure of somatic mutation load, we divided the number of variants in each motif occurrence (or matching motif-free region) by their length. Motifs and motif-free regions overlapping with RepeatMasker annotations were excluded also from this analysis.

Comparison of errors and variants between motifs and motif-free regions. Illumina errors, polymorphic variants, fixed variants and somatic mutations are all rare events, resulting in a large portion of motif occurrences and motif-free regions with rates exactly equal to 0. As a result the distributions of rates among motifs, as well as among motif-free regions, have an excess of 0-valued observations - corresponding to regions without errors/variants. More specifically, each distribution comprises a spike at 0, together with a distribution on strictly positive values. Thus, to compare rates between motif occurrences and matching motif-free regions, we employed a two-part test (*93*, *94*) that contrasts both the heights of spikes at 0 and the distributions of positive values. Notably, PacBio errors are more abundant and only a small portion of regions have rates equal to 0. However, for consistency we analyzed them employing the same two-part test. The compound null hypothesis is that both the spike at 0 (proportion of 0 rates) and the distribution on positive values (continuous component on non-0 rates) are the same in the two groups, versus the two-sided alternative that either or both differ between the groups. We considered the two-part statistic $V^2 = B^2 + T^2$, where B^2 is the continuity-corrected binomial test statistic (contrasting the proportions of 0 rates) and T^2 is the square of the t-test statistic (contrasting the non-0 rates). P-values were generated approximating the distribution of the test statistic V^2 under the null hypothesis with a $\chi^2(2)$, in order to overcome the computational burden of estimating its distribution using permutations. For several cases, we also computed p-values based on 10,000 random permutations and obtained almost indistinguishable results. For

robustness, each test was repeated 10 times, using separate sets of randomly generated matching motif-free regions, and significance was assessed based on the maximum p-value (maximum p-values ≤0.10 are coded by standard stars-and-dots representation in Fig. 4). When the two parts of the test statistic suggested opposite directions in the comparison between motifs and motif-free regions (e.g., motifs showed an increased proportion of 0 rates, but also an increased mean for non-0 rates), we marked the corresponding result as inconclusive. White cells in Fig. 4 represent non-significant cases, inconclusive cases, and cases with insufficient number of events (sum of events in all motif occurrences < 20).

In addition to the tests, we also computed rate fold differences (the numbers in Fig. 4) as follows. For each motif type, we considered the whole portion of the genome covered by its occurrences. For comparison, we considered the portion of the genome covered by *all* 100-bp motif-free windows (note: *not* matching motif-free regions). Error rates (PacBio and Illumina data) were estimated dividing total number of errors by total number of bases sequenced in the considered portion of the genome. Mutation rates (divergence, diversity and TCGA data) were estimated dividing the total number of variants by the total length of the considered portion of the genome. Rate fold differences were then computed, for each motif type and each error/variant type, as motif rate over motif-free rate if the former is larger, and motif-free rate over motif rate otherwise.

Data and Code availability. All scripts are available in public repository https://bitbucket.org/makova-lab/kinetics_wmm. Readers are encouraged to download the latest versions of the scripts directly from the BitBucket repository. The data are available at Extended Data Files 1 and 2.

SUPPLEMENTARY NOTES

SUPPLEMENTARY NOTE 1. Estimation of falsely reported errors due to misalignment.

Measurement of sequencing error rates from aligned reads can be affected by misalignments. When the target genome contains identical or nearly identical regions, an aligner may map some reads to the wrong locations. A true variant at one position can, after being mismapped, appear to be a sequencing error at a different position. Additionally, variants in very close proximity can be misreported (for example two nearby indels may be reported as one indel and a few mismatches). Because non-B DNA contains motifs, i.e. repetitive regions by definition, such misalignments might be particularly common in them. To evaluate the prevalence of "false errors" induced by misalignment, we performed alignment on simulated sequencing data and compared error calls to the known truth.

Methods. A haploid mock genome was constructed consisting of non-B DNA motifs of six different types (motifs intersecting RepeatMasker intervals were removed; motifs longer than 100 bp were shortened to their central 100 bp). The mock genome copied the motifs and 99-bp flanks on each side from hg19. Motifs were separated by runs of 100 'N's, unless flanking regions overlapped.

100 bp reads were randomly sampled to 60x coverage from the mock genome, with a simplified error model similar to the one described by Schirmer and colleagues(*95*). Each base had a 0.2% chance of a mismatch, a $5x10^{-6}$ chance of a single base insertion, and a $5x10^{-6}$ chance of a single base deletion. The ground truth of induced errors was recorded.

Reads were aligned to the mock genome with bwa mem(*96*) (default settings). Naive Variant Caller(*88*) was used to detect errors in the aligned reads, as well as in the ground truth, and the two sets of calls were compared. Only reads mapping to reverse strand were used. "False errors", i.e. those induced by misalignment, are the calls present in the aligned reads but absent from the ground truth. The following table reports, for each motif type and error type, the number of false errors observed in this experiment. Rates are reported relative to (a) the number of positions in the mock genome, and (b) the number of bases in simulated reads. The rate per read base is estimated from the rate per genome position and the 60X depth of simulated reads.

Our results indicate that misalignments account for a small fraction of the observed Illumina errors. For nearly all motifs and error types, the false error rate is below 10% of the Illumina error rate. The only exceptions are insertions in Z-DNA and in A-phased repeats (13.5% and 12% of Illumina errors, respectively, likely due to the small number of data points).

Feature	Event type	False error events	Bases considered	False rate per genome position	False rate per read base	Illumina error rate per read base on reverse strand only (for comparison*)
APhased repeats	mismatches	300	311218	9.64E-04	1.61E-05	3.22E-03
APhased repeats	deletions	32	311218	1.03E-04	1.71E-06	3.25E-05
APhased repeats	insertions	25	311218	8.03E-05	1.34E-06	1.08E-05
Direct repeats	mismatches	6424	577789	1.11E-02	1.85E-04	1.50E-02
Direct repeats	deletions	68	577789	1.18E-04	1.96E-06	1.09E-04
Direct repeats	insertions	54	577789	9.35E-05	1.56E-06	6.72E-05
GQuadPlus	mismatches	134	173697	7.71E-04	1.29E-05	7.73E-03
GQuadPlus	deletions	20	173697	1.15E-04	1.92E-06	1.29E-04
GQuadPlus	insertions	11	173697	6.33E-05	1.06E-06	9.60E-05
GQuadMinus	mismatches	151	170158	8.87E-04	1.48E-05	7.62E-03
GQuadMinus	deletions	21	170158	1.23E-04	2.06E-06	1.31E-04
GQuadMinus	insertions	14	170158	8.23E-05	1.37E-06	6.45E-05
Inverted repeats	mismatches	4672	3942305	1.19E-03	1.98E-05	3.72E-03
Inverted repeats	deletions	363	3942305	9.21E-05	1.53E-06	4.61E-05
Inverted repeats	insertions	236	3942305	5.99E-05	9.98E-07	1.54E-05
Mirror repeats	mismatches	1151	946446	1.22E-03	2.03E-05	4.31E-03
Mirror repeats	deletions	93	946446	9.83E-05	1.64E-06	4.04E-05
Mirror repeats	insertions	44	946446	4.65E-05	7.75E-07	1.60E-05
ZDNA motifs	mismatches	63	34921	1.80E-03	3.01E-05	6.00E-03
ZDNA motifs	deletions	0	34921	0.00E+00	0.00E+00	2.23E-06
ZDNA motifs	insertions	5	34921	1.43E-04	2.39E-06	1.77E-05

*From Extended Data File 1, both Read 1 and Read 2 considered here, see also Supplementary Note 3.

SUPPLEMENTARY NOTE 2. Impact of different aligners on calling sequencing errors.

The five datasets we used for the analyses presented in Fig. 4 were generated by different projects employing different aligners; namely: blasr(*97*) (for SMRT errors), novoalign(*98*) (for Illumina errors), lastz(*99*) (for diversity; 1000 Genome Project), multiz(*100*) (for divergence; human-orangutan alignments), and bwa(*96*) (for cancer somatic mutations; TCGA). The exact placements of variants can differ between sequence aligners and between aligner parameterizations, especially in repeat regions. Consequently, measures of event rates for various event types at various locations may be aligner-dependent.

To evaluate the extent to which the results in Fig. 4 could be affected by the aligners used, we applied five different aligners commonly used for Illumina reads to the Illumina sequencing data (*101*) used for the Illumina results presented in Fig. 4 and Supplementary Note 3.

Methods. The pipeline we employed to call Illumina sequencing errors was repeated with five aligners (novoalign, bwa mem, bowtie2(*102*), last(*103*), and stampy(*104*)). To limit computation time, we restricted attention to motifs annotated on chromosome 1 (motifs intersecting RepeatMasker intervals and variants compared to hg19 were removed; motifs longer than 100 bp were shortened to their central 100 bp) and the matching motif-free regions across the genome (but only for one of the 10 sets of motif-free regions we generated for the analyses in Fig. 4). Reads 1 and 2 previously identified by novoalign alignment to the whole genome as aligning to the reverse complement sequences of the motifs considered (to study errors on the newly synthesized strand using motifs as a template) were aligned again using default (or typical) parameters. The alignment target for each read was restricted to that read's previously-identified chromosome. Alignments from each aligner were then independently processed using the Illumina sequencing errors pipeline.

The following Table reports fold-differences in Illumina sequencing error rates between motifs (on the non-repetitive portion of chromosome 1) and matched motif-free regions (see Methods). Error rates are derived using five different aligners — novoalign, bwa, bowtie2, last, and stampy — separately for (**A**) mismatches, (**B**) deletions and (**C**) insertions. Red/blue is used to indicate higher/lower rates in motifs than in motif-free regions. Motifs types considered, with corresponding sample sizes in parentheses, are: A-phased repeats (n=945), direct repeats (n=1,050), inverted repeats (n=15,819), mirror repeats (n=1,186), Z-DNA (n=241), G-quadruplexes on the reference strand (G4+; n=645), and G-quadruplexes on the reverse complement strand (G4-; n=645). NA: cell value not computable (lacking error calls in either motifs or motif-free regions for the cell). The full data underlying this Table is available in the code repository as Extended Data File 2.xlsx.

Our results indicate that, while the aligners did induce some differences in the fold-differences computed for various event rates, the overall trends were very similar across aligners. Note that our exercise is not informative for deletions and insertions in Z-DNA motifs and for insertions in G4-; NAs in the Table. This is due to lack of data. Also, results for insertions in direct repeats and G4+ show some instability; see red/blue in the Table. This is

not concerning for direct repeats because all of our short-read sequencing rate fold-differences for insertions in them were not significant (Fig. S17). The results for insertions in G4+ should be interpreted with caution; the corresponding fold-differences for short-read data in Fig. 4C are marginally significant.

Mismatches	Aphased	Direct	Inverted	Mirror	ZDNA	G4+	G4-
bwa	-1.33	2.33	-1.03	1.12	1.82	1.83	1.83
bowtie	-1.26	3.43	-1.03	1.19	1.80	1.87	2.26
last	-1.44	2.22	-1.02	1.16	1.65	1.88	2.01
novoalign	-1.34	2.25	-1.02	1.14	1.74	1.86	1.86
stampy	-1.40	3.14	-1.03	1.16	1.69	1.89	2.06

Deletions	Aphased	Direct	Inverted	Mirror	ZDNA	G4+	G4-
bwa	-1.17	1.72	2.07	1.38	NA	5.70	2.68
bowtie	-1.89	1.94	2.11	1.62	NA	6.34	2.07
last	-1.52	-1.28	2.03	1.56	NA	5.26	3.61
novoalign	-1.58	1.81	1.94	1.57	NA	5.31	3.34
stampy	-1.56	2.27	1.98	1.25	NA	6.09	3.79

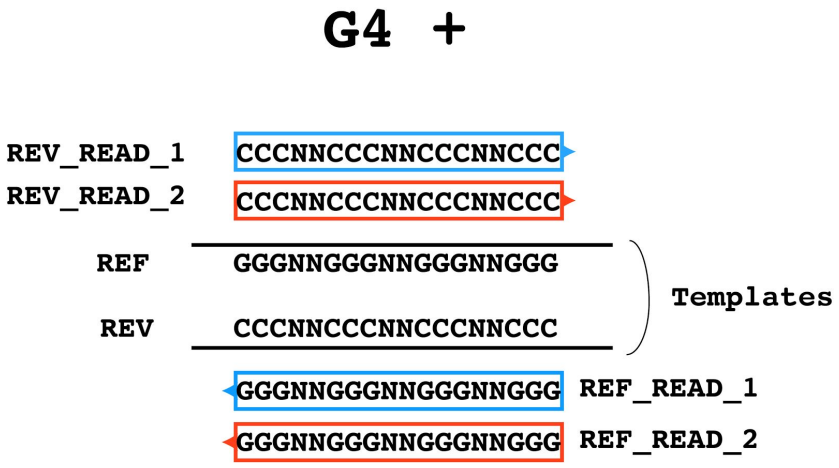
Insertions	Aphased	Direct	Inverted	Mirror	ZDNA	G4+	G4-
bwa	-8.52	-1.98	-1.28	1.35	NA	NA	1.12
bowtie	-3.19	3.68	-1.08	1.65	NA	2.11	1.73
last	-23.15	-1.06	-1.24	1.60	NA	NA	NA
novoalign	-12.08	-2.88	-1.50	1.44	NA	NA	NA
stampy	-2.57	4.36	-1.28	1.67	NA	1.93	2.60

SUPPLEMENTARY NOTE 3. Strand orientation and error calculations for Illumina paired-end sequencing.

Illumina paired-end sequenced reads, which correspond to the newly synthesized strand, can be split according to their mapping orientation (reference or reverse) with respect to the reference genome. This allows us to investigate strand-specific effects of non-B DNA motifs. Moreover, on average, from a pair, Read 1, which is synthesized first, has fewer errors per base than Read 2. Thus, Reads 1 and 2 should be analyzed separately.

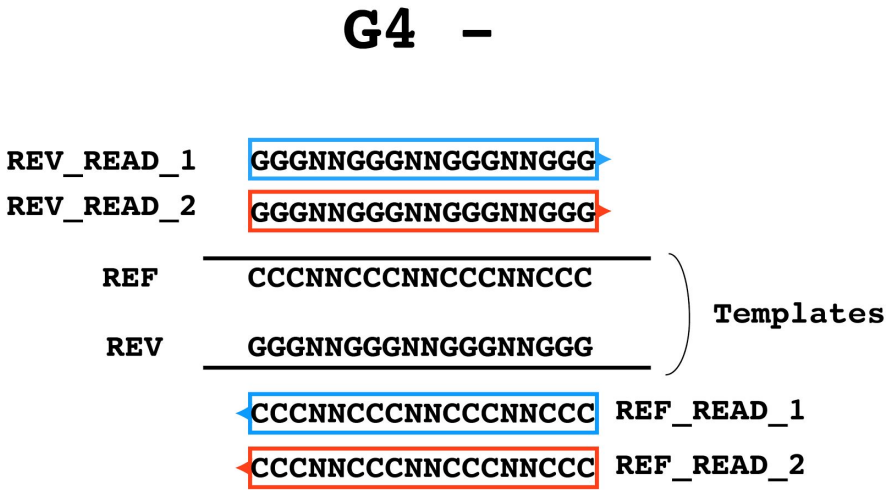
For G4+, the G-quadruplex motif is located on the reference strand, and this sequence can be read by Illumina instrument by four different types of reads (Fig. A). To measure the effects of a G4-containing template on the newly synthesized C-rich strand, we should analyze C-rich reverse reads (REV_READ_1 and REV_READ_2), i.e. the reads mapping to the reverse complement of the G-quadruplex. REV_READ_1 is expected to contain fewer errors than REV_READ_2. To measure the effects of C-rich template on the newly synthesized G-rich strand, we should analyze G-rich REF_READ_1 and REF_READ_2, both mapping to the reference. Here again REF_READ_1 is expected to contain fewer errors than REF_READ_2.

Figure A



Similarly for G4-, the G-quadruplex motif is located on the reverse strand, and this sequence can be read by Illumina instrument by four different types of reads (Fig. B). To measure the effects of a G-rich template on the newly synthesized C-rich strand, we should analyze C-rich reference reads (REF_READ_1 and REF_READ_2), i.e. the reads mapping to the reference G4-. REF_READ_1 is expected to contain fewer errors than REF_READ_2. To measure the effects of C-rich template on the newly synthesized G-rich strand, we should analyze G-rich REV_READ_1 and REV_READ_2, both mapping to the reference. Here again REV_READ_1 is expected to contain fewer errors than REV_READ_2.

Figure B



Using the methods used to generate Fig. 4 in the main text, we analyzed the error rates for the four read types for G4+ and for the four read types for G4- (Fig. C, see legend of Fig. 4 in the main text; raw data can be found in Extended Data File 1).

Figure C

Mismatches	A-phased repeats	Direct repeats	Inverted repeats	Mirror repeats	Z-DNA	G4+ motifs	G4- motifs
REF_READ_1	1.27 ***	4.07 ****	1.07 ****	1.18 **	1.52 *	2.35 ****	2.71 ****
REF_READ_2	1.19 ****	2.76 ****	1.06 ****	1.06 *	1.33 ***	1.97 ****	1.35 ****
REV_READ_1	1.20 **	4.29 ****	1.05 ***	1.16 **	1.53 *	2.71 ****	2.73 ****
REV_READ_2	1.13 **	2.75 ****	1.07 ****	1.09 *	1.35 ****	1.36 ****	2.00 ****
REV_READ_1 + REV_READ_2	1.15 ****	3.09 ****	1.06 ****	1.10 **	1.39 ***	1.62 ****	2.20 ****

Fold-differences and significance assessments (see legend of Fig. 4 in the main text) are consistent across rows. However, both G4+ and G4- motifs have elevated error rates, and we cannot determine a clear strand bias for this increase. One potential explanation of increased Illumina errors at G4- motifs is propagation of errors at G4+ motifs via bridge amplification. On average, we detected more errors on READ_2 than on READ_1 (Extended Data File 1). For this reason, we are only presenting READ_1 results in the main text.

SUPPLEMENTARY NOTE 4. Evaluating the potential impact of increased Illumina error rate on high-frequency variant calling for non-B DNA motifs.

In our analysis of sequencing errors (Fig. 4), we demonstrated that Illumina sequencing accuracy is affected by non-B DNA motifs. Indeed, several non-B DNA motif types lead to increased Illumina errors. Thus, because the 1000 Genomes Project employs Illumina technology, the impact of sequencing errors on the detection of variants might be higher in non-B DNA motifs than in motif-free regions.

In order to quantitate the impact of increased error rate on high-frequency variant calling (global frequency>5%) in the 1000 Genomes Project, we estimated Illumina error rates in the different types of non-B DNA motifs. For each motif type, we computed the total Illumina error rate per read base f_M as the sum of mismatches, insertions and deletions, divided by the total number of bases sequenced (see Supplementary Note 3), combining all occurrences of the considered motif. We then modeled the Illumina sequencing error process at a nucleotide belonging to the motif type M as a Bernoulli trial $X_M \sim B(1, f_M)$. Similarly, we computed the total error rate per nucleotide f_C for motif-free regions (controls), and we modeled the baseline Illumina error process as $X_C \sim B(1, f_C)$.

Assume that, for each of the 5,008 individual haploid genomes in the 1000 Genomes Project (corresponding to 2,504 individuals), each site is sequenced exactly once (i.e. exactly one read maps to it) for each individual. Then, in each genome, the variants per nucleotide observed because of Illumina error are $X_M \sim B(1, f_M)$ and $X_C \sim B(1, f_C)$, for nucleotides belonging to motifs of type M and motif-free regions, respectively. Note that this assumption is very conservative, since all individuals are actually sequenced at a depth higher than 4x. If we further assume that sequencing errors for different haploid genomes are independent, the number of haplotypes (out of 5,008) with a variant on a single site due to Illumina sequencing error is $V_M \sim B(5,008; f_M)$ for motif type M and $V_C \sim B(5,008; f_C)$ for motif-free regions. In the worst case scenario, in which all errors at the same site produce the same variant, the corresponding probabilities that detection of a high-frequency variant is due solely to sequencing error can be computed as $P(V_M \geq 251)$ and $P(V_C \geq 251)$, where 251 = 5,008×5%, because we use variants with global frequency above 5%.

The following table reports, for each motif type and for motif-free regions, the estimated Illumina total error rate, the probability that a high-frequency variant in the 1000 Genome Project data is due to sequencing error, the number of high-frequency variants actually detected in these data, and the expected number of variants due to sequencing errors (assuming that variants are independent). Both the probability and the number of expected variants are extremely low and thus cannot explain our results.

	Illumina total error rate	Prob variant due to errors	1000 Genome variants	Expected variants due to errors
Motif-free	2.01×10^{-3}	6.87×10^{-251}	230,300	1.58×10^{-245}
A-phased rep	1.61×10^{-3}	4.06×10^{-274}	672	2.73×10^{-271}

Direct rep	7.57×10^{-3}	1.00×10^{-117}	1,544	1.54×10^{-114}
Inverted rep	1.89×10^{-3}	4.74×10^{-257}	9,306	4.41×10^{-253}
Mirror rep	2.17×10^{-3}	9.63×10^{-243}	2,339	2.25×10^{-239}
Z-DNA	2.97×10^{-3}	3.34×10^{-210}	176	5.88×10^{-208}
G4	4.50×10^{-3}	3.91×10^{-168}	1,112	4.35×10^{-165}

A previous study of HiSeq data(*95*) demonstrated that the occurrence Illumina errors depend on sequence context. In particular, substitutions were shown to depend on the 3-mers preceding them. The datasets analyzed in that study had an overall substitution error rate of 3.15×10^{-3} (2.1×10^{-3} and 4.2×10^{-3} errors per base in read 1 and 2, respectively), but the 3-mer "GGG" alone accounted for up to 17% of all substitutions (that is, up to 11 times more than expected by chance). This systematic bias can have a strong impact on the variants observed in G-quadruplexes, which contain many occurrences of the "GGG" 3-mer.

To quantitate this impact, we employ again the binomial model introduced above restricting attention to substitutions only and assuming, as a worst case scenario, that all 1,112 variants observed in G-quadruplexes occurred in nucleotides immediately following an occurrence of the "GGG" 3-mer. In this positions Illumina error rate can be as high as 0.035, and the expected number of variants due to sequencing errors is equal to 2.73×10^{-5}. Although the number of expected variants is higher following this 3-mer, it is still very low (less than 1 variant is expected to be observed because of errors) and cannot explain our results.

The conservative calculations presented in this Supplementary note show that (i) high-frequency variants detected in the 1000 Genome Project data, be those within non-B DNA motifs or motif free regions, are extremely unlikely to be caused solely by Illumina sequencing error; and (ii) high-frequency substitutions detected in the same data within G-quadruplexes are also extremely unlikely to be caused by systematic sequence context biases in Illumina sequencing errors.

SUPPLEMENTARY NOTE 5. *De novo* mutations from deCODE genetics Iceland trios.

We utilized 108,778 recently published *de novo* mutations discovered in 1,548 Icelandic trios (*68*) to test whether these mutations are enriched in non-B DNA motifs. The data were downloaded from the study PRJEB21300 in the European Nucleotide Archive (https://www.ebi.ac.uk/ena/data/view/PRJEB21300). We recovered 1,548 vcf files (one per trio) and all variants passing all quality filters (FILTER ID = PASS). A total of 108,778 variant coordinates were translated from hg38 to hg19 using the Lift-Over tool in Galaxy (*88*). Finally, using the methods resulting in Fig.4 in the main text, we obtained fold-differences in *de novo* mutations between non-B DNA motifs and 10 control sets. Because of the small number of *de novo* mutations, we were able to compute meaningful fold-differences only for mismatches, and also for these, the mutation sample sizes were too small to reach statistical significance. Fold-differences for *de novo* mismatches are reported below (the raw data are in the Extended Data File 1).

	A-phased	Direct	Inverted	Mirror	Z-DNA	G4
Mismatches	-1.19	-1.43	1.03	-1.02	NA	1.13

SUPPLEMENTARY TEXT

Interval-Wise Testing: statistical details.

The IWT is a novel inferential procedure for functional data(*55*) that performs a global two-sample test on the whole domain of the curves being compared, and simultaneously detects locations where the difference between the two samples of curves is significant. The IWT was developed in order to overcome weaknesses of the two testing procedures(*105*, *106*) previously proposed in the FDA literature to deal with the same inferential problem. These procedures both required an initial discretization step: the Interval Testing Procedure (ITP)(*105*) was based on a basis expansion of the curves, while the procedure developed in(*106*) utilized an a priori partition of the curve domain in smaller intervals. Different discretization choices in this initial step can affect test results and conclusions. Notably, despite this issue, the ITP was successfully employed in(*107*) to characterize the genomic landscape surrounding endogenous retrovirus locations in human and mouse. The IWT does not require discretization; it operates directly on the original curves, providing more reliable results. Moreover, being a non-parametric permutation test, it can be employed even if the data distribution is skewed, which is the case in our application to IPD values (Fig. S2). Here we used an extended version of the IWT specifically designed for "Omics" data applications(*108*). This extension outputs both the locations and the scales that lead to rejecting a null hypothesis. In addition, it allows the user to select among different test statistics that highlight complementary characteristics of the curve distributions.

Let $IPD_{f,i}(t)\ \ i = 1, ..., n_f$ be the IPD curves in the n_f motif-containing windows (features), and $IPD_{c,i}(t)\ \ i = 1, ..., n_c$ the IPD curves in the n_c motif-free windows (controls). Each curve is defined in the interval $I = (-50, 50)$ (0 representing the center of the motif for motif-containing windows) and comprises 100 values corresponding to the 100 nucleotides where the IPD is measured. Missing IPD measurements are treated as gaps in the curves. We treat $IPD_{f,i}(t)\ i = 1, ..., n_f$ and $IPD_{c,i}(t)\ \ i = 1, ..., n_c$ as two random samples from two independent random functions, and test the null hypothesis H_0^I that the two random functions have the same distribution over the whole interval I, versus the alternative H_1^I that they have different distributions. When we detect significant differences between the two IPD curve distributions (i.e. when we reject the null hypothesis), we aim to identify the portions of the curves (locations) where these differences occurs. Moreover, we want to select the lengths $s = |S|$ (scales) of the subintervals $S = (t_a, t_b) \subseteq I$ where these differences are strong enough to be detected by restricting the null hypothesis to S (indicated as H_0^S).

For each subinterval $S \subseteq I$, we define the mean test statistic as

$$T_{mean}(S) = \frac{1}{|S|}\int_S \left(\overline{IPD_f}(t) - \overline{IPD_c}(t)\right)^2 dt$$

where $\overline{IPD_f}(t) = \frac{1}{n_f}\sum_{i=1}^{n_f} IPD_{f,i}(t)$ and $\overline{IPD_c}(t) = \frac{1}{n_c}\sum_{i=1}^{n_c} IPD_{c,i}(t)$ are the sample means of the IPD curves in the two groups. Similarly, we define the median and the multi-quantile test statistics as

$$T_{median}(S) = \frac{1}{|S|}\int_S \left(IPD_f^{0.50}(t) - IPD_c^{0.50}(t)\right)^2 dt$$

and

$$T_{multi-quantile}(S) = \sum_{q\in Q}\frac{1}{|S|}\int_S \left(IPD_f^{q}(t) - IPD_c^{q}(t)\right)^2 dt$$

where, for every $t \in I$, $IPD_f^q(t)$ and $IPD_c^q(t)$ are the quantiles of order q of the IPD curves in the n_f motif-containing windows and n_c motif-free windows, respectively, and Q is a given set of probabilities. Different statistics allow us to focus on different characteristic of the curve distributions. In particular, if the set Q spans a large portion of $[0,1]$, the multi-quantile statistic is very effective in leveraging information on the whole curve distributions. For example, we can use the quartiles ($Q = \{0.25, 0.50, 0.75\}$) to capture differences in the central part of the distribution, or we can add smaller and larger quantiles ($Q = \{0.05, 0.25, 0.50, 0.75, 0.95\}$) to capture also differences in the tails.

Given a choice of the test statistic T, the first step of the IWT is a functional permutation test for the hypothesis H_0^S versus H_1^S on every subinterval $S = (t_a, t_b) \subseteq I$ and every complementary interval $S = I \setminus (t_a, t_b)$. In particular, we estimate the empirical distribution of the test statistic T under H_0^S conditionally to the data, by evaluating $T(S)$ for all possible permutations of the $n_f + n_c$ observed curves, and we compute the test p-value p^S as the proportion of permutations that lead to a test statistic greater than or equal to the one evaluated on the original data (two-sided test, note that the test statistic is non-negative). The second step of the IWT generates an adjusted p-value curve $\tilde{p}(t)$, defined in each $t \in I$ as

$$\tilde{p}(t) = \sup_{S \ni t} p^S.$$

This multiple testing correction controls the interval-wise error rate; that is, $\tilde{p}(t)$ controls the probability of rejecting the null hypothesis H_0^S on every interval $S \subseteq I$ where it is true (see details in(*109*)). Finally, we identify locations with a significant difference in motif vs motif-free windows by selecting all $t \in I$ such that $\tilde{p}(t) \leq \alpha$, where α is the desired significance level.

In order to detect the scales at which the differences in IPD are significant, the extended IWT evaluates multiple scales, generating an adjusted p-value curve $\tilde{p}_s(t)$ for each scale $s \leq |I|$. In particular, for each fixed s, $\tilde{p}_s(t)$ considers only the subintervals $S \subseteq I$ of length $|S| \leq s$ and thus controls the interval-wise error rate on all intervals of length at most s. As a consequence, the extended IWT identifies significant locations for all possible scales s (i.e. the points $t \in I$ such that $\tilde{p}_s(t) \leq \alpha$).

SUPPLEMENTARY TABLES

Table S1. Nucleotides annotated in non-B DNA motifs in the human genome.
The number of nucleotides annotated for each motif type according to the non-B DB(*110*) (and according to STR-FM(*111*) for STRs). Nucleotides may be annotated as part of one or more motifs.

Motifs	**Sequence Definition according to non-B DB**	**Counts**
Direct repeats	10-50 nt repeated within 5 nt spacer	42,300,423
Mirror repeats	10-100 nt mirrored within 100 nt spacer	77,078,820
Inverted repeats	10-100 nt with reverse complement within 100 nt spacer	133,278,477
A-phased repeats	3 or more A-tracts (3-5 As) 10 nt on center each; Spacers between equal sized A-tracts must contain some non As	10,504,652
Z-DNA motis	G followed by Y (C or T) for at least 10 nt; One strand must be alternating Gs	6,700,444
G-quadruplex motifs	4 or more G-tracts (3-7 Gs) separated by 1-7 nt spacers; Preference for short spacers with Cs and/or Ts	10,102,937
STRs	Tandem repeats of 1-4 base pairs per motif	187,657,110

Table S2. Tested non-B DNA motifs.

The last two columns represent the sample size for each motif type on each strand.

Motif	Structure	On both strands	Number of windows with annotation	Number of windows after filtering for overlaps	Number of windows with IPD on reference strand	Number of windows with IPD on reverse strand
A-Phased repeats	slipped-strand	yes	404,289	26,218	26,142	26,143
Direct repeats	slipped-strand	yes	1,501,567	34,778	34,582	34,594
Inverted repeats	cruciform	yes	6,365,102	470,135	468,525	468,520
Mirror repeats	H-DNA	yes	1,895,543	43,053	39,919	39,932
Z-DNA motifs	Z-DNA	yes	412,600	6,229	6,207	6,209
G-quadruplex motifs	G-quad	no	181,230 (+) 180,213 (-)	13,125 (+) 12,971 (-)	13,049(+) 12,876 (-)	13,046 (+) 12,885 (-)

Table S3. Tested STRs.

We studied the motif-specific effect of STRs by collapsing all alignable motifs using the method described in Table S9. Motifs with less than 15 windows having IPD on reference or reverse strand (in gray) were not analyzed. The last two columns represent the sample size for each motif in the two strands.

Motif	Number of windows with annotations	Number of windows after filtering for overlaps	Number of windows with IPD on reference strand	Number of windows with IPD on reverse strand
$(A)_n$	6,727,074	583,681	581,804	581,800
$(C)_n$	1,263,551	135,124	134,603	134,600
$(G)_n$	1,263,833	135,109	134,571	134,564
$(T)_n$	6,758,517	585,904	583,991	584,027
$(AC)_n$	1,281,488	127,385	126,947	126,947
$(AG)_n$	1,607,242	166,884	166,312	166,296
$(AT)_n$	2,107,265	117,575	117,242	117,244
$(CG)_n$	60,759	6,427	6,378	6,381
$(CT)_n$	1,608,739	167,349	166,754	166,749
$(GT)_n$	1,291,081	128,972	128,520	128,525
$(AAC)_n$	68,259	3,919	3,909	3,909
$(AAG)_n$	86,740	7,042	7,020	7,019
$(AAT)_n$	167,160	9,230	9,209	9,209
$(ACC)_n$	114,798	32,880	32,736	32,739
$(ACG)_n$	592	18	70	71
$(ACT)_n$	16,998	1,404	1,402	1,402
$(AGC)_n$	62,444	7,454	7,421	7,421
$(AGG)_n$	84,147	7,740	7,706	7,712
$(AGT)_n$	16,875	1,408	1,405	1,405
$(ATC)_n$	53,402	3,839	3,829	3,829
$(ATG)_n$	52,944	3,871	3,858	3,858
$(ATT)_n$	166,990	9,078	9,050	9,058
$(CCG)_n$	9,297	413	411	410
$(CCT)_n$	84,257	7,743	7,702	7,705
$(CGG)_n$	9,424	427	426	425
$(CGT)_n$	591	71	71	71
$(CTG)_n$	63,715	7,687	7,660	7,655

$(CTT)_n$	86,491	7,246	7,229	7,226
$(GGT)_n$	114,492	32,743	32,576	32,576
$(GTT)_n$	68,914	3,793	3,779	3,782
$(AAAC)_n$	41,472	1,579	1,573	1,571
$(AAAG)_n$	31,680	1,096	1,093	1,093
$(AAAT)_n$	61,622	2,904	2,891	2,894
$(AACC)_n$	1,735	122	122	122
$(AACG)_n$	25	3	3	2
$(AACT)_n$	453	37	37	37
$(AAGC)_n$	1,444	107	107	107
$(AAGG)_n$	11,944	443	440	440
$(AAGT)_n$	776	74	74	74
$(AATC)_n$	2,633	246	246	246
$(AATG)_n$	15,190	1,347	1,345	1,345
$(AATT)_n$	8,704	62	62	62
$(ACAG)_n$	2,849	232	232	232
$(ACAT)_n$	6,599	155	154	154
$(ACCC)_n$	3,090	144	144	144
$(ACCG)_n$	23	2	2	2
$(ACCT)_n$	870	59	59	59
$(ACGG)_n$	70	2	2	2
$(ACTC)_n$	2,884	247	246	246
$(ACTG)_n$	945	98	98	98
$(ACTT)_n$	749	54	54	54
$(AGAT)_n$	5,583	104	104	104
$(AGCC)_n$	2,522	229	229	229
$(AGCG)_n$	186	10	10	10
$(AGCT)_n$	673	11	11	11
$(AGGC)_n$	5,237	325	323	323
$(AGGG)_n$	10,619	368	367	366
$(AGGT)_n$	901	67	66	66
$(AGTC)_n$	916	76	75	75
$(AGTG)_n$	2,841	253	249	249
$(AGTT)_n$	403	35	35	35

$(ATCC)_n$	5,940	217	216	216
$(ATCT)_n$	5,575	112	112	112
$(ATGC)_n$	2,277	21	21	21
$(ATGG)_n$	6,009	179	179	179
$(ATGT)_n$	6,755	172	171	172
$(ATTC)_n$	15,055	1,434	1,433	1,431
$(ATTG)_n$	2,708	242	242	242
$(ATTT)_n$	62,007	2,933	2,927	2,924
$(CCCG)_n$	840	18	18	18
$(CCCT)_n$	10,734	376	375	375
$(CCGG)_n$	348	6	6	6
$(CCGT)_n$	44	1	1	1
$(CCTG)_n$	5,267	341	340	341
$(CCTT)_n$	11,829	444	444	444
$(CGCT)_n$	156	10	10	10
$(CGGG)_n$	804	23	23	23
$(CGGT)_n$	17	2	2	2
$(CGTT)_n$	34	2	2	2
$(CTGG)_n$	2,311	224	223	223
$(CTGT)_n$	2,787	185	184	184
$(CTTG)_n$	1,412	120	120	120
$(CTTT)_n$	32,220	1,136	1,131	1,131
$(GGGT)_n$	3,260	170	170	170
$(GGTT)_n$	1,750	154	154	154
$(GTTT)_n$	41,692	1,533	1,529	1,528

Table S4. Non-B DNA potential (in addition to slipped-strand structures) for microsatellite sequences.

Hairpin (self-complementary)	**H-DNA (poly Pur or Poly Pyr)**	**Z-DNA (Pur-Pyr)**
$(AT)_n$ (Ref (*112*); a cruciform)	$(A)_n$ (Ref (*112*); also form A tract/bent)	$(AC)_n$ (Ref (*112*))
$(AAT)_n$ (predicted from sequence)	$(C)_n$ (Ref (*112*))	$(CG)_n$ (Ref (*112*))
$(ACT)_n$ (predicted from sequence)	$(G)_n$ (Ref (*112*); also form A tract/bent)	$(GT)_n$ (Ref (*112*))
$(AGC)_n$ (Ref (*113*))	$(T)_n$ (Ref (*112*))	
$(AGG)_n$ (Ref (*114*))	$(AG)_n$ (Ref (*112*))	
$(AGT)_n$ (predicted from sequence)	$(CT)_n$ (Ref (*112*))	
$(ATC)_n$ (predicted from sequence)	$(AAG)_n$ (Ref (*112*))	
$(ATG)_n$ (predicted from sequence)	$(CCT)_n$ (predicted from sequence)	
$(ATT)_n$ (Ref (*115*))	$(CTT)_n$ (Ref (*112*))	
$(CCG)_n$ (Ref (*113*))		
$(CGG)_n$ (Ref (*113*))		
$(CTG)_n$ (Ref (*113*))		

Table S5. Measures of G-quadruplex stability and structure determined by Circular Dichroism for the ten most common G-quadruplex motifs in the genome.
G1 through G10 indicate, in the order of frequency in the genome, the ten most common G-quadruplex motif types in our annotations (G1 -- the most common, G2 the next most common, etc.). The last column reports the number of occurrences of each motif type after filtering out the ones completely lacking IPD values and the distribution of the mean IPD. Cyan indicates intra-stranded G-quadruplexes, while orange indicates inter-stranded ones. “Intra” -- intramolecular, “bimol” -- bimolecular, “paral” -- parallel structures, “anti” -- antiparallel structures.

Sequence	T_m [°C]	Molecularity	Max delta epsilon	Strand orientation	Mean IPD (5th, 25th, 50th, 75th, 95th quantiles)
G1 GGGGTGGGGGGAGGGGGGAGGG	74.3	intra	248	paral + anti	0.91 1.07 1.19 1.33 1.60 (2,962 occurrences)
G2 GGGAGGGAGGTGGGGGGG	64.8	bimol	298	paral	0.86 0.98 1.06 1.18 1.36 (540 occurrences)
G3 GGGGTCGGGGGAGGGGGGAGGG	74.8	intra	216	paral + anti	0.75 0.84 0.91 0.98 1.14 (440 occurrences)
G4 GGGGTGGGGGGAGTGGGGAGGG	69.0	intra	209	paral + anti	0.74 0.83 0.90 0.99 1.13 (312 occurrences)
G5 GGGAGGGAGGGAGGGAGGG	69.0	bimol 2 types	300	paral	0.84 0.99 1.15 1.29 1.62 (287 occurrences)
G6 GGGAGGGAGGTGGGGGGG	68.0	bimol + higher	300	paral	0.81 0.97 1.06 1.16 1.36 (148 occurrences)
G7 GGGTGGAGGGTGGGAGGAGGG	61.5	bimol 2 types	282	paral	0.83 0.92 1.00 1.08 1.28 (262 occurrences)
G8 GGGGTTGGGGGAGGGGGGAGGG	73.2	intra	211	paral + anti	0.78 0.85 0.93 1.01 1.21 (189 occurrences)
G9 GGGGTGGGGGGAGGGGGAGGG	71.9	intra	281	paral + anti	0.93 1.17 1.38 1.66 2.09 (181 occurrences)
G10 GGGGTGGGGGGAGCGGGGAGGG	68.5	intra	216	paral + anti	0.82 0.91 1.00 1.07 1.32 (177 occurrences)

Table S6. Measures of $(GGT)_n$ motif stability and structure determined by Circular Dichroism.

Cyan indicates intra-stranded structures, while orange indicates inter-stranded ones. See other abbreviations explained in the previous table.

Sequence	T_m [°C]	Molecularity	Max delta epsilon	Strand orientation
$(GGT)_4$ GGTGGTGGTGGT	48.0	tetra	184	paral + anti
$(GGT)_5$ GGTGGTGGTGGT GGT	45.2	bimol	138	paral
$(GGT)_6$ GGTGGTGGTGGT GGTGGT	39.0	bimol + intra	117	paral + anti

Table S7. Sample size (the number of motifs) for computing and testing fold differences in the rates of sequencing errors and mutations.

G4+ and G4- are combined for mutations (diversity, divergence, and TCGA data).

Motifs	**Sample size for sequencing errors**	**Sample size for mutations**
A-phased repeats	10,895	12,108
Direct repeats	12,423	13,704
Inverted repeats	168,191	187,200
Mirror repeats	13,185	14,700
Z-DNA motifs	2,764	3,103
G-quadruplexes on the reference (G4+)	5,938	12,984
G-quadruplexes on the reverse complement (G4-)	5,696	

Table S8. Complete data for fold differences in error / mutation rates when the reverse complement strand is used as a template and motifs are annotated on it.
Red indicates increase, while blue decrease, over motif-free regions. Cells shaded in gray have lack of data (fewer than 10 error or mutation events). The rates and test p-values are provided in the Extended Data File 1. Illumina errors are reported for REF_READ_1 only.

Mismatches	Aphased	Direct	Inverted	Mirror	ZDNA	G4+	G4-
Pacbio	-1.0278	1.0929	1.0047	1.0188	-1.1512	1.0586	1.7827
Illumina	-1.2702	4.0690	-1.0663	1.1756	1.5201	2.3524	2.7149
Divergence	-1.1128	-1.5926	-1.0758	-1.0188	1.7808	1.1523	
1000G	-1.1283	-1.0311	-1.0483	1.0154	1.9419	1.3021	
TCGA	-1.4172	-2.2654	-1.1787	-1.0425	2.5497	1.1289	

Deletions	Aphased	Direct	Inverted	Mirror	ZDNA	G4+	G4-
Pacbio	-1.0438	1.0233	-1.0003	1.0011	-1.1736	1.1023	1.4886
Illumina	1.1493	2.9510	1.3488	1.1964	-2.1392	4.3729	2.9630
Divergence	1.0735	-1.4416	1.0877	1.1999	2.4162	1.1489	
1000G	1.3433	-1.4093	1.3850	1.1191	-1.9008	1.3114	
TCGA	-1.0053	-1.1967	1.3948	1.3143	1.3329	2.5263	

Insertions	Aphased	Direct	Inverted	Mirror	ZDNA	G4+	G4-
Pacbio	-1.0100	-1.0159	-1.0198	-1.0096	1.1698	-1.2335	-1.0220
Illumina	-1.3183	3.3542	1.4116	2.2492	-2.2368	4.6688	6.0163
Divergence	-1.3450	11.6942	1.5363	1.4272	2.0465	3.7499	
1000G	-1.2118	-1.5089	1.3144	1.3570	4.6138	3.0588	
TCGA	1.0267	-1.5515	-1.0779	1.2762	2.1560	1.6138	

Table S9. STR aligning and collapsing: an example.
The five STRs shown in the table are aligned and collapsed to allow correct motif alignment, and presented as the motif $(ACTT)_n$. A capitalized nucleotide indicates the center of the STR, while bracketed nucleotides show near-central positions chosen to align the motifs.

Motif	**STR**	**Aligned microsatellite**
$(ACTT)_2$	acttActt	actt[A]ctt
$(CTTA)_3$	cttactTactta	cttactT[a]ctta
$(TTAC)_3$	ttacttActtac	ttactt[A]cttac
$(TACT)_5$	tacttacttaCttacttact	tacttactt[a]Cttacttact
$(ACTT)_4$	acttacttActtactt	acttactt[A]cttactt

SUPPLEMENTARY FIGURES

Figure S1. Window centering of motifs with an even or odd number of nucleotides. Each box is a nucleotide. The red box/line represent the motif and window centers.

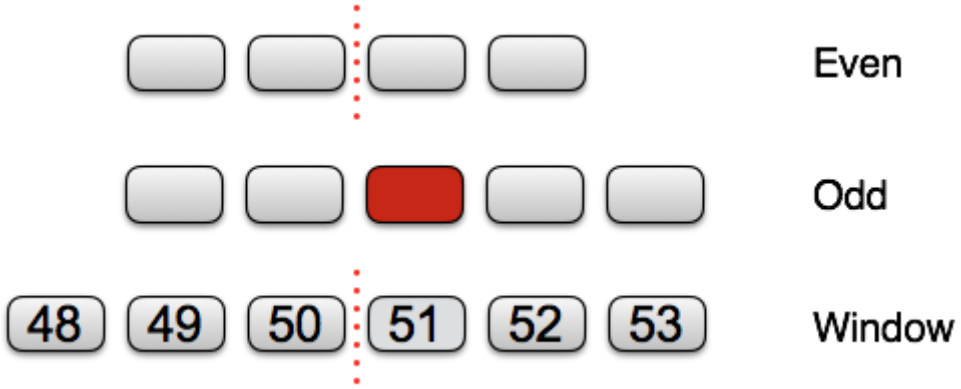
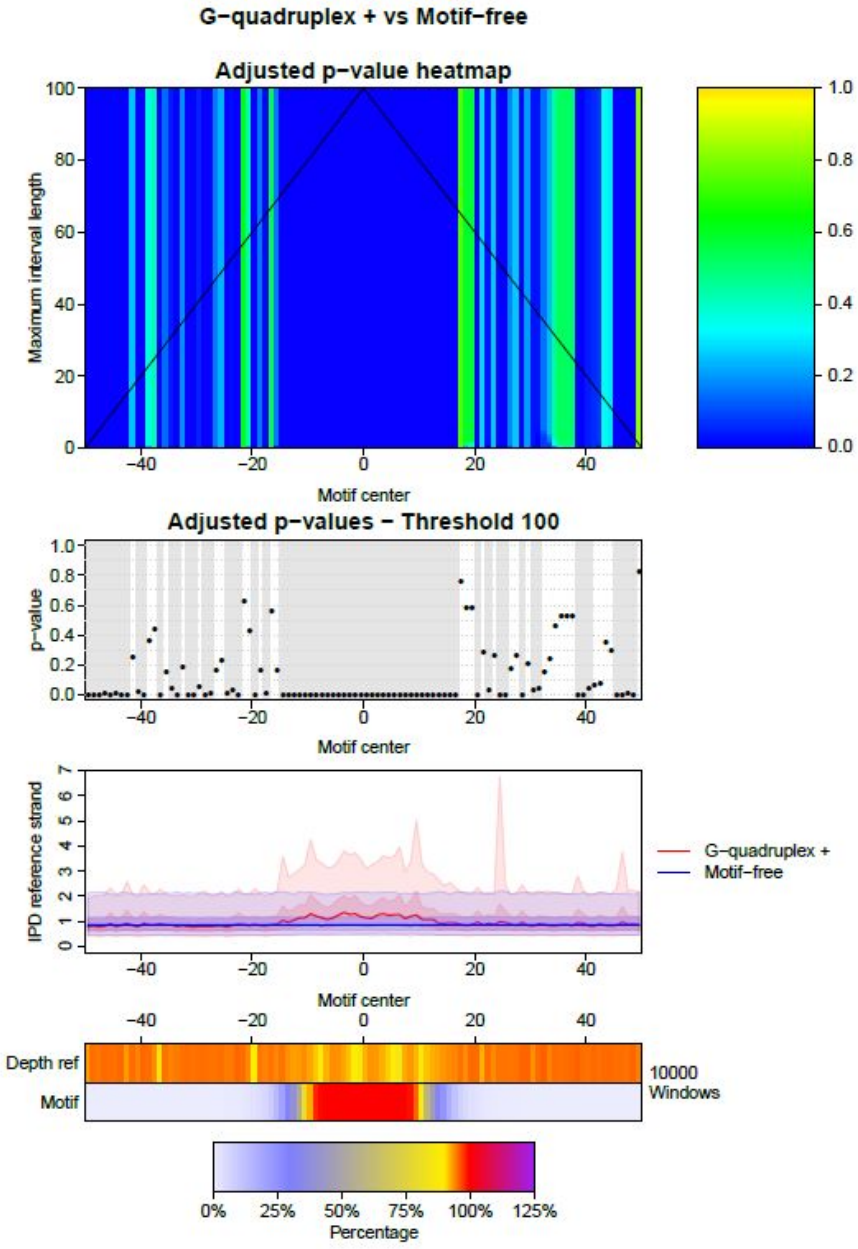


Figure S2. An example of detailed results of Interval-Wise Testing.
Results of IWT using multi-quantile statistic and a random subsample of 10,000 windows for the comparisons **A** G-quadruplex motifs on reference strand vs. motif-free windows. **B** $(AGC)_n$ vs. motif-free windows. The heatmap at the top shows the p-value curves produced by the IWT for every possible scale. The x axis indicates the positions in the 100-bp window. The y axis indicates the scale at which the test is performed, from the 1-bp scale (bottom row of the heatmap, maximum interval length=1) to the maximum possible scale of 100-bp (top row of the heatmap, maximum interval length=100). Blue corresponds to low p-values. The central plot shows the p-value curve at scale 100-bp, with gray areas highlighting significant positions (p-values≤0.05). The plot and heatmap at the bottom show the distribution of IPD values (see caption of Fig. 2A).

A



B

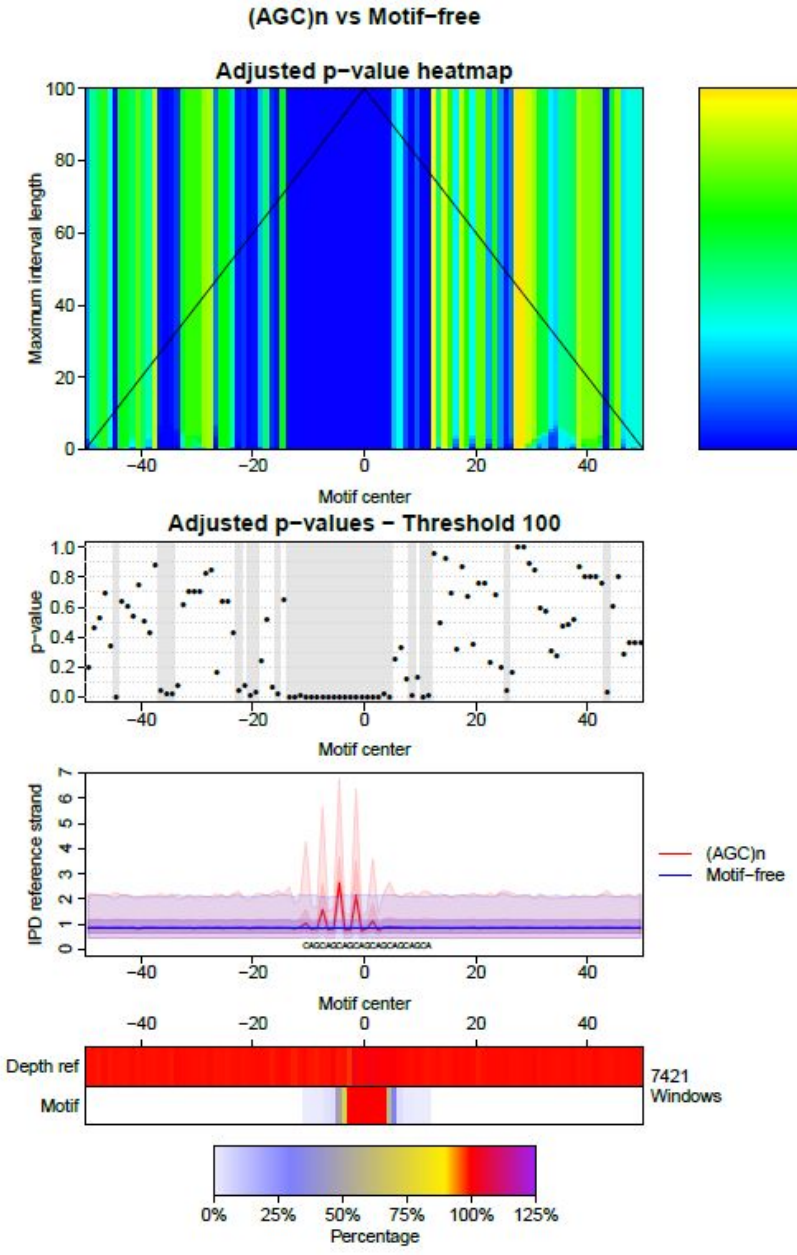
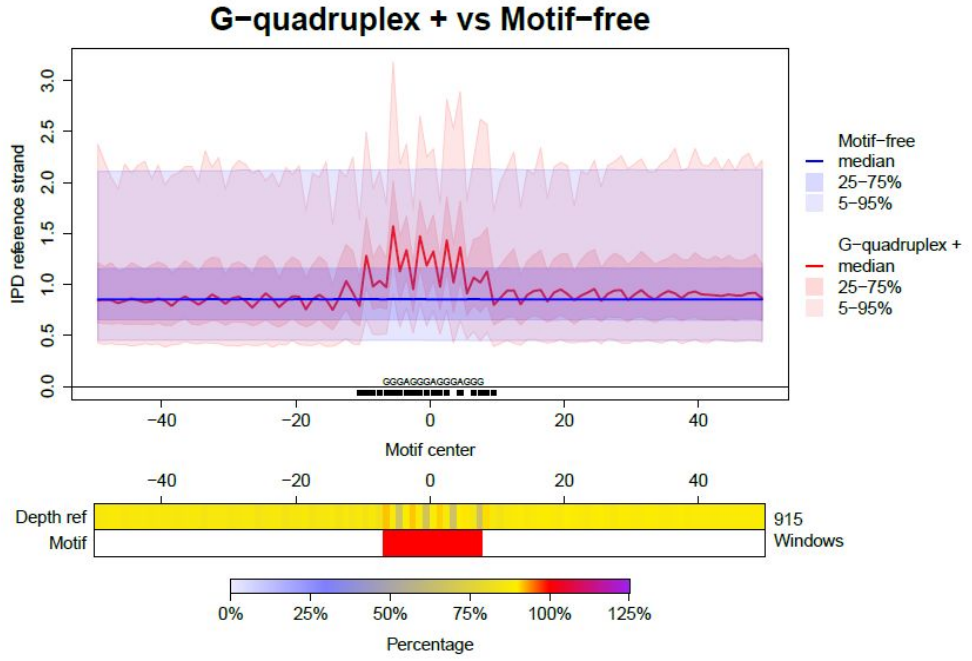

(AGC)n vs Motif−free
Adjusted p−value heatmap
Maximum interval length
Motif center
Adjusted p−values − Threshold 100
p−value
Motif center
IPD reference strand
(AGC)n
Motif−free
Motif center
Depth ref
Motif
7421
Windows
0%
25%
50%
75%
100%
125%
Percentage

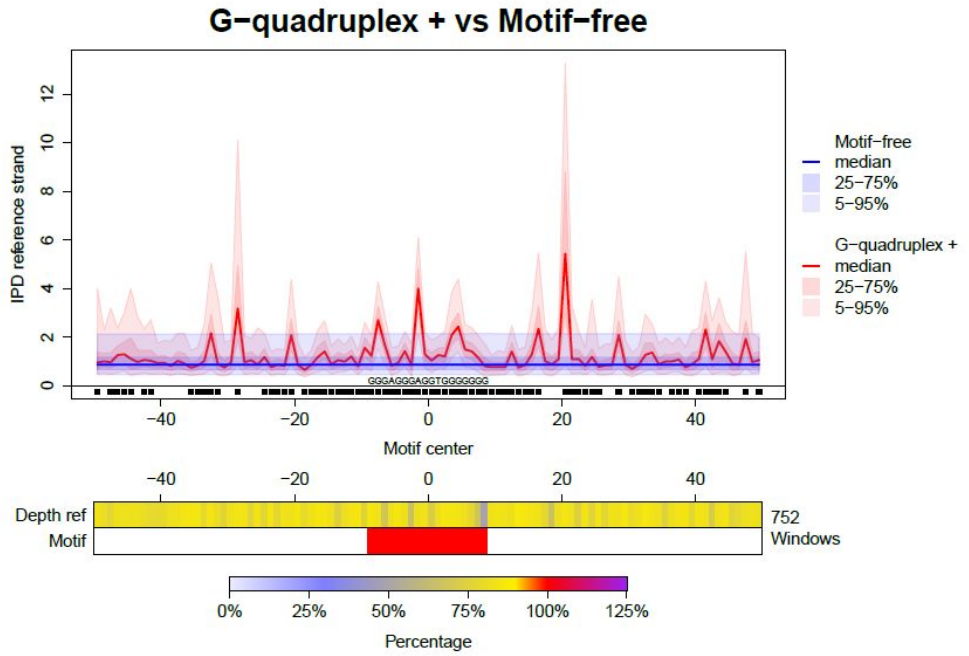
Figure S3. Different shapes of IPD curve distributions among different G-quadruplex motifs.

The analysis dividing G4 motifs based on their motifs was performed on the full data set of >300,000 G4 motifs, allowing overlaps between motifs of the same and different types - we do not have enough data to perform such an analysis for our non-overlapping data set of 26,000 motifs. The results still confirm elevated IPDs at G4s demonstrating that filtering for overlapping annotations does not affect our main results. **A** $GGGA_{3-5}G_3$ motifs only have the central elevation and lack the 3' spike. **B** $GGGA_2GGT_1G_{7-8}$ and **C** $G_3T_1G_2A_1G_3T_1G_3A_1G_2A_1G_3$ present only spikes in 5', 3' and overlapping the motif. **D** $G_4TN_1G_5A_1G_6A_1G_3$, **E** $G_4T_1G_5A_2G_6A_1G_3$, **F** $G_4T_1G_6A_{1-2}G_5A_1G_3$, **G** $G_4T_1G_6AGN_1G_4A_1G_3$, **H** $G_4T_1G_6A_1G_5A_{1-2}G_3$ and **I** $G_4T_1G_6AT_1G_5A_1G_3$ all have a central elevation surrounded by spikes as well as the 3' spike. Finally, **J** $GGGT_3GGG_1$ shows a series of periodic spikes, similar to the pattern observed at many microsatellites. This suggests that the last motif actually folds into a slipped structure and not into a G-quadruplex. See the legend of Fig. 2A.

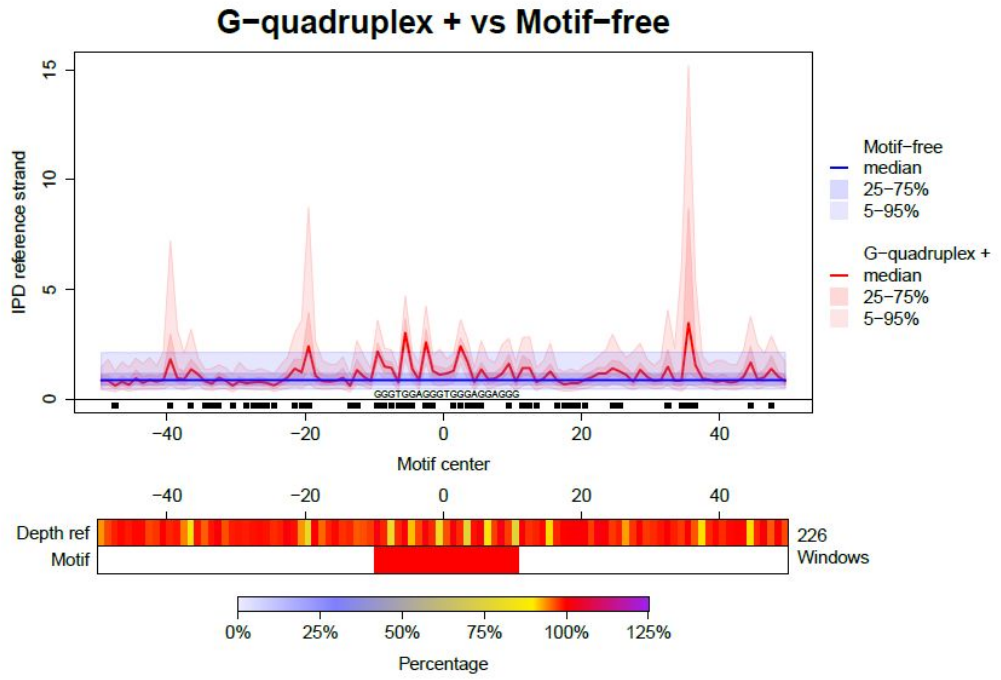
A



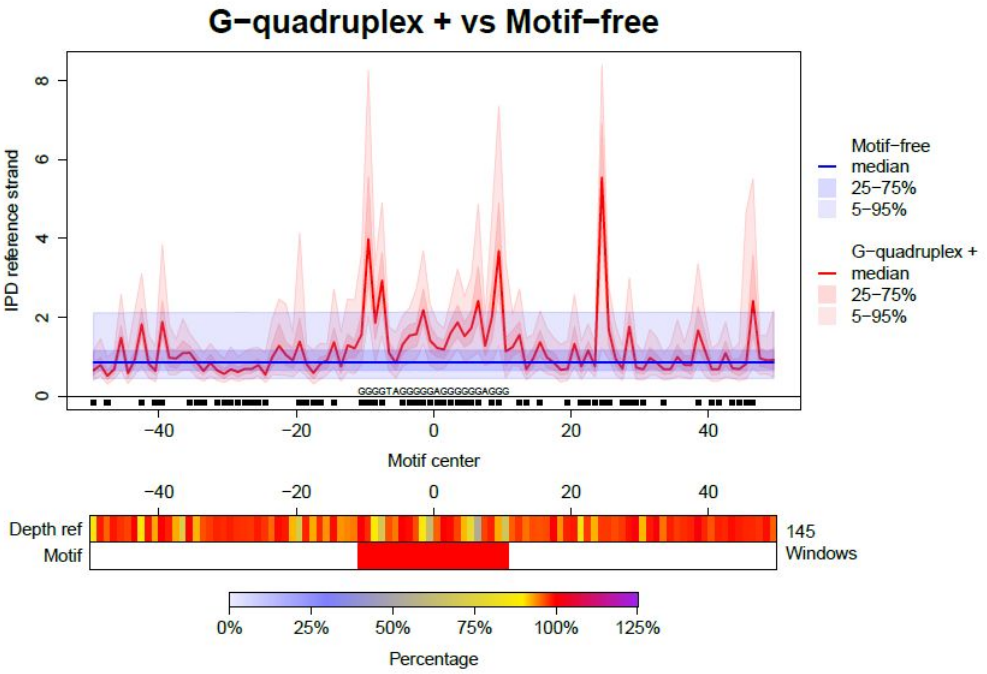
B


G−quadruplex + vs Motif−free
IPD reference strand
0
2
4
6
8
10
12
GGGAGGGAGGTGGGGGGG
−40
−20
0
20
40
Motif center
Depth ref
Motif
752
Windows
0%
25%
50%
75%
100%
125%
Percentage
Motif−free
median
25−75%
5−95%
G−quadruplex +
median
25−75%
5−95%

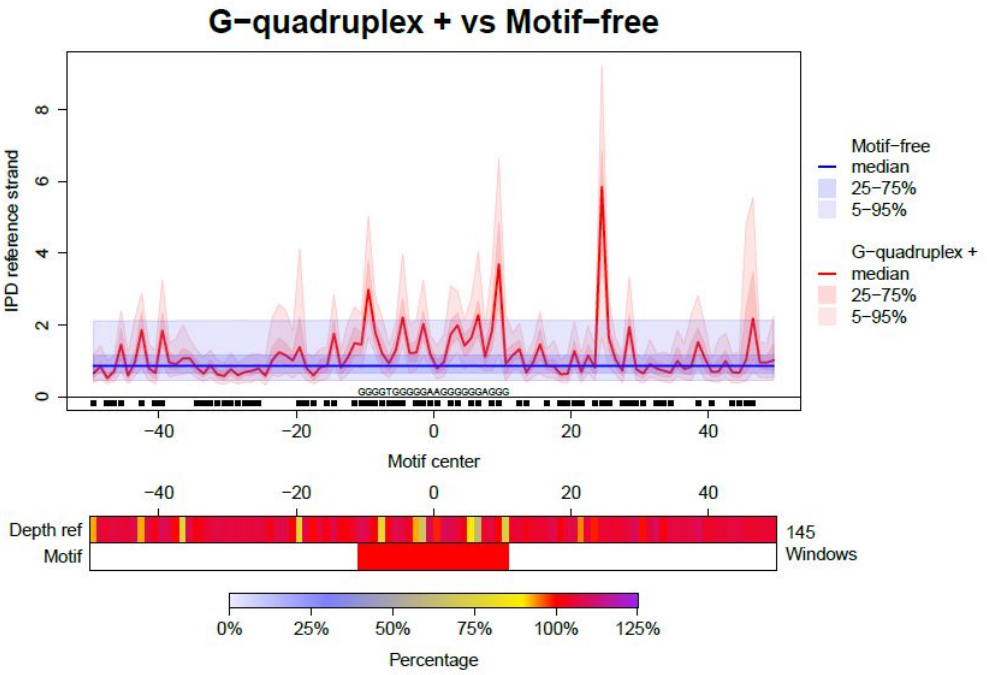
C


G−quadruplex + vs Motif−free
IPD reference strand
0
5
10
15
GGGTGGAGGGTGGGAGGAGGG
−40
−20
0
20
40
Motif center
Depth ref
Motif
226
Windows
0%
25%
50%
75%
100%
125%
Percentage
Motif−free
median
25−75%
5−95%
G−quadruplex +
median
25−75%
5−95%

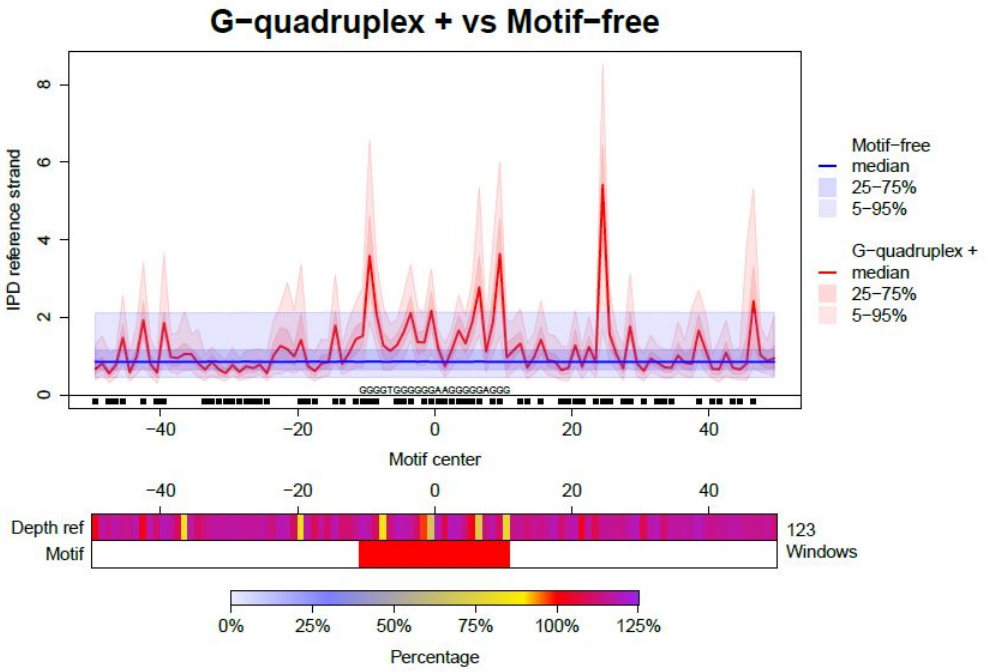
D


G−quadruplex + vs Motif−free
IPD reference strand
Motif−free
median
25−75%
5−95%
G−quadruplex +
median
25−75%
5−95%
GGGGTAGGGGGAGGGGGGAGGG
−40
−20
0
20
40
Motif center
Depth ref
Motif
145
Windows
0%
25%
50%
75%
100%
125%
Percentage

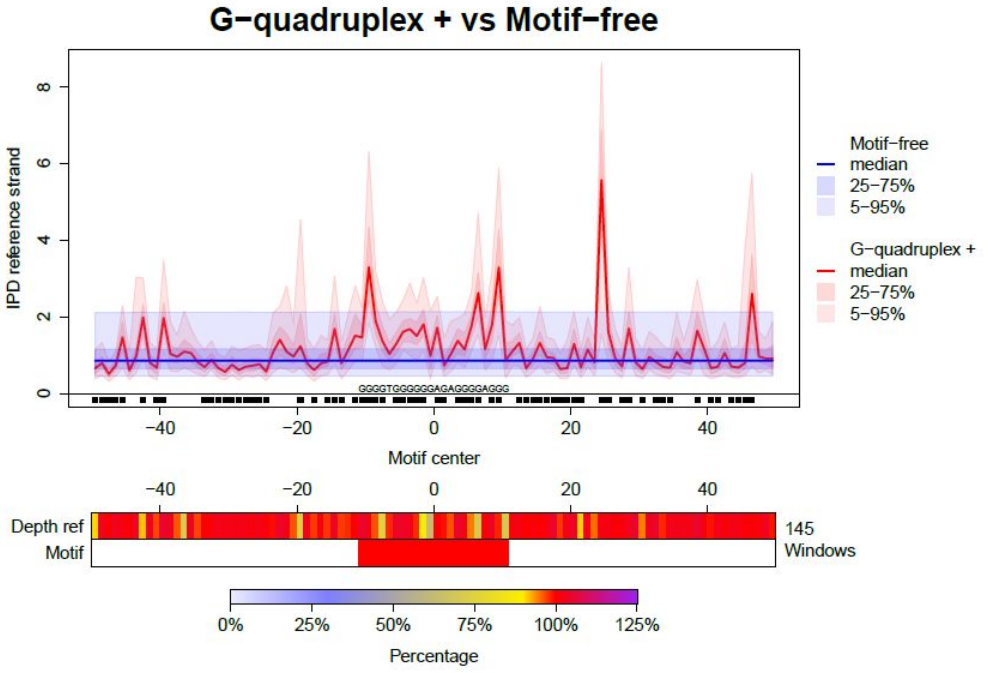
E


G−quadruplex + vs Motif−free
IPD reference strand
Motif−free
median
25−75%
5−95%
G−quadruplex +
median
25−75%
5−95%
GGGGTGGGGGAAGGGGGGAGGG
−40
−20
0
20
40
Motif center
Depth ref
Motif
145
Windows
0%
25%
50%
75%
100%
125%
Percentage

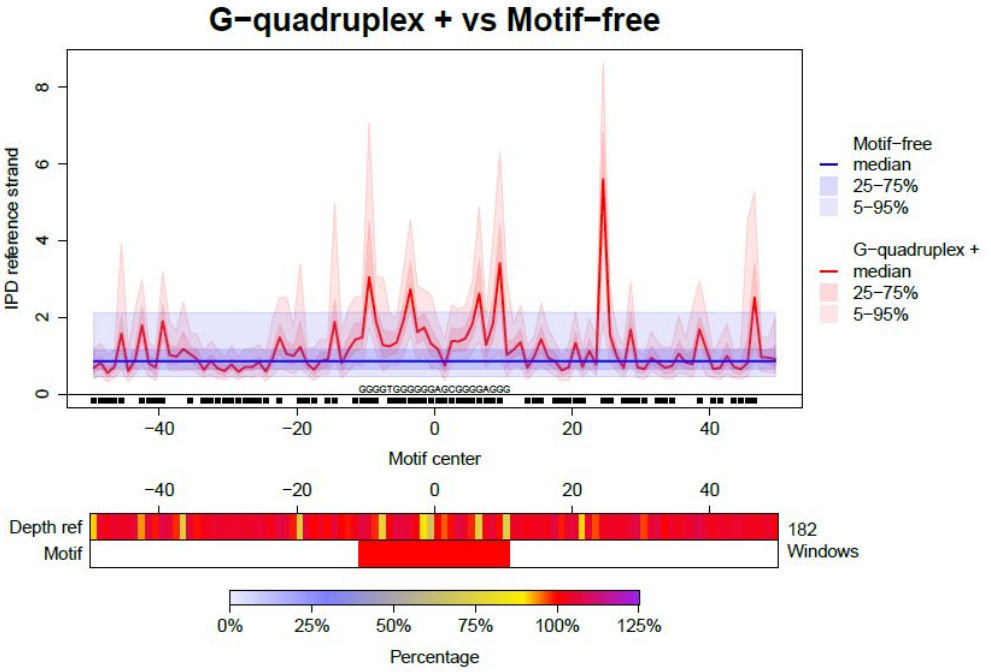
F


G−quadruplex + vs Motif−free
IPD reference strand
Motif−free
median
25−75%
5−95%
G−quadruplex +
median
25−75%
5−95%
GGGGTGGGGGGAAGGGGGAGGG
−40
−20
0
20
40
Motif center
Depth ref
Motif
123
Windows
0%
25%
50%
75%
100%
125%
Percentage

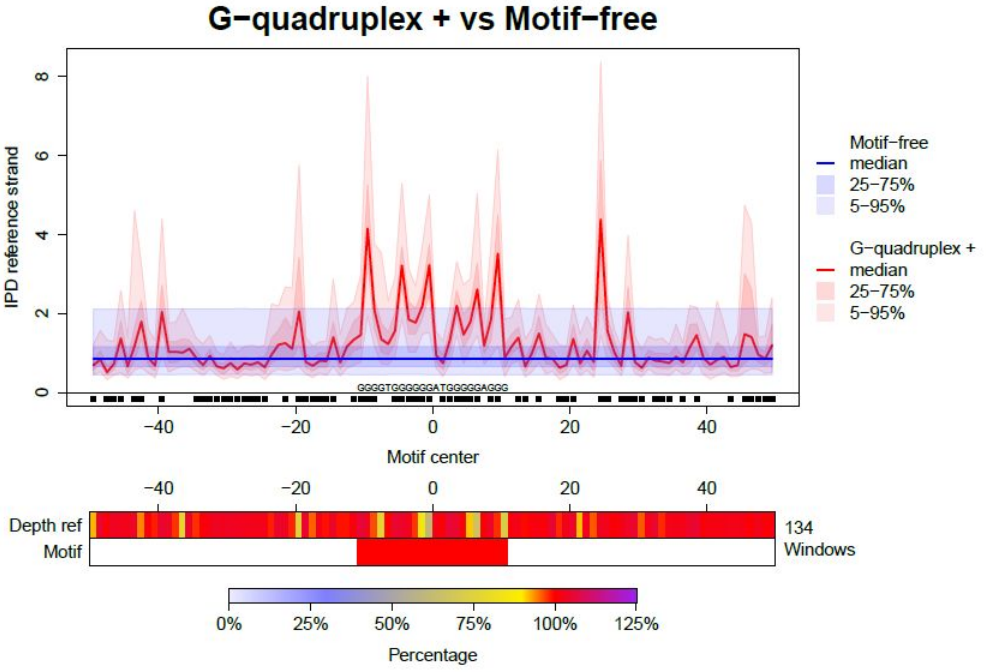
G


G−quadruplex + vs Motif−free
IPD reference strand
Motif−free
median
25−75%
5−95%
G−quadruplex +
median
25−75%
5−95%
GGGGTGGGGGGAGAGGGGAGGG
−40
−20
0
20
40
Motif center
Depth ref
Motif
145
Windows
0%
25%
50%
75%
100%
125%
Percentage

H


G−quadruplex + vs Motif−free
IPD reference strand
GGGGTGGGGGGAGCGGGGAGGG
Motif center
Depth ref
Motif
182
Windows
0%
25%
50%
75%
100%
125%
Percentage
Motif−free
median
25−75%
5−95%
G−quadruplex +
median
25−75%
5−95%

I


G−quadruplex + vs Motif−free
IPD reference strand
GGGGTGGGGGGATGGGGGAGGG
Motif center
Depth ref
Motif
134
Windows
0%
25%
50%
75%
100%
125%
Percentage
Motif−free
median
25−75%
5−95%
G−quadruplex +
median
25−75%
5−95%

J

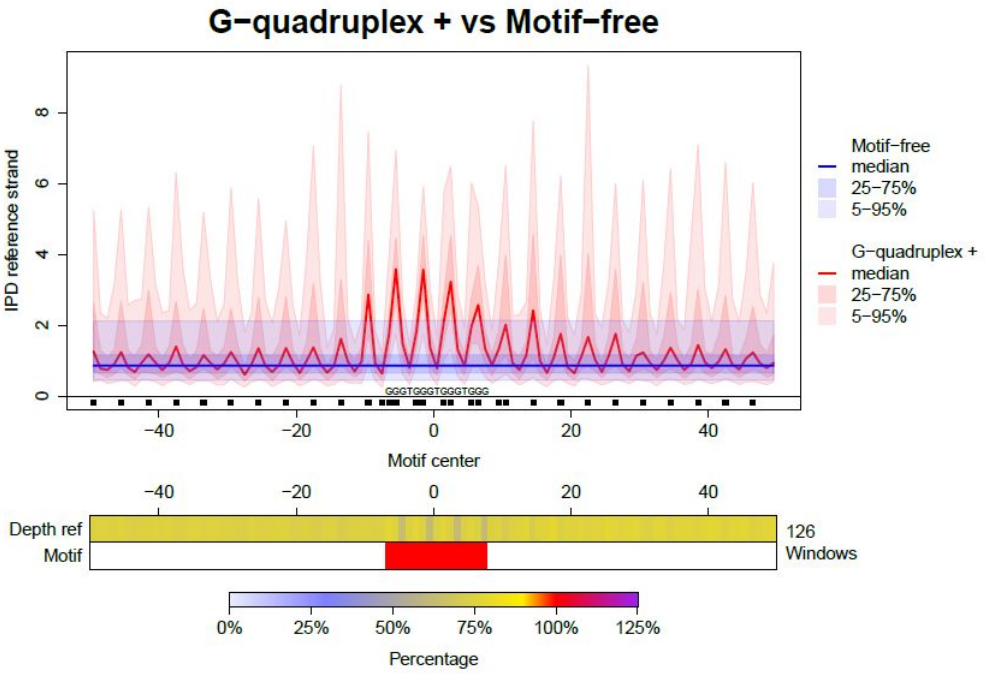
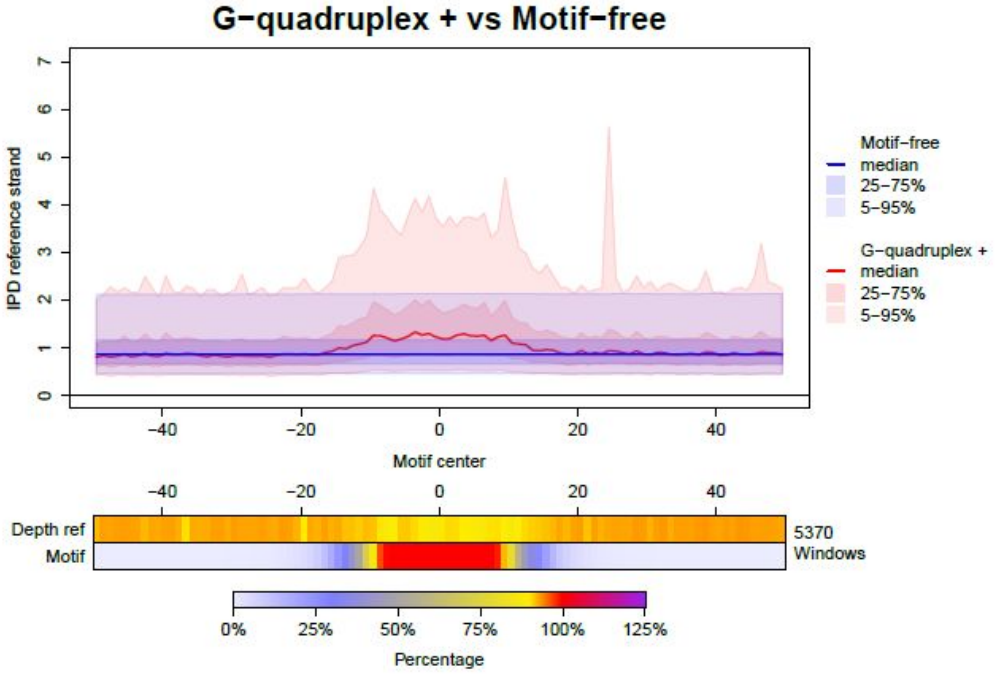

G−quadruplex + vs Motif−free
IPD reference strand
0
2
4
6
8
GGGTGGGTGGGTGGG
−40
−20
0
20
40
Motif center
Depth ref
Motif
126
Windows
0%
25%
50%
75%
100%
125%
Percentage
Motif−free
median
25−75%
5−95%
G−quadruplex +
median
25−75%
5−95%

Figure S4. IPD curve distribution for G-quadruplexes identified by in vitro ion concentration manipulations.

A The IPD profile for G4+ on the reference strand (computed on 5,370 windows) is very similar to the one obtained considering all G4+ motifs (13,049 windows; see top panel of Fig. 2A). **B** The IPD profile for G4- on the reference strand (computed on 5,463 windows) is very similar to the mirror image of the one obtained considering all G4+ motifs on the reverse complement strand (13,046 windows; see bottom panel of Fig. 2A). No statistical test was performed. Additional details on various elements of these graphical representations can be found in the legend of Fig. 2A.

A



B

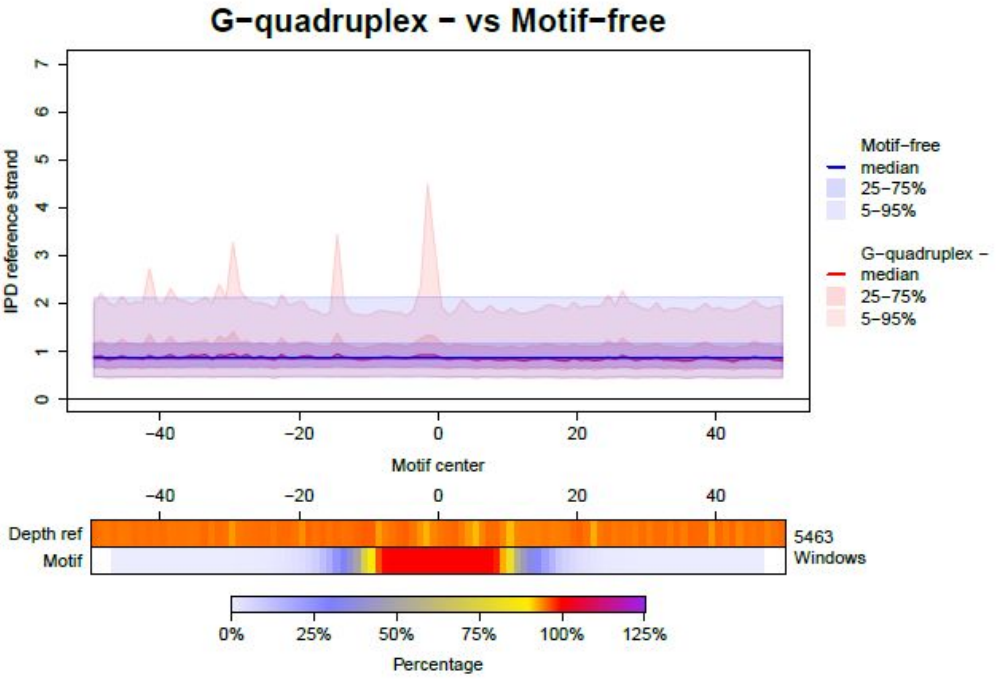
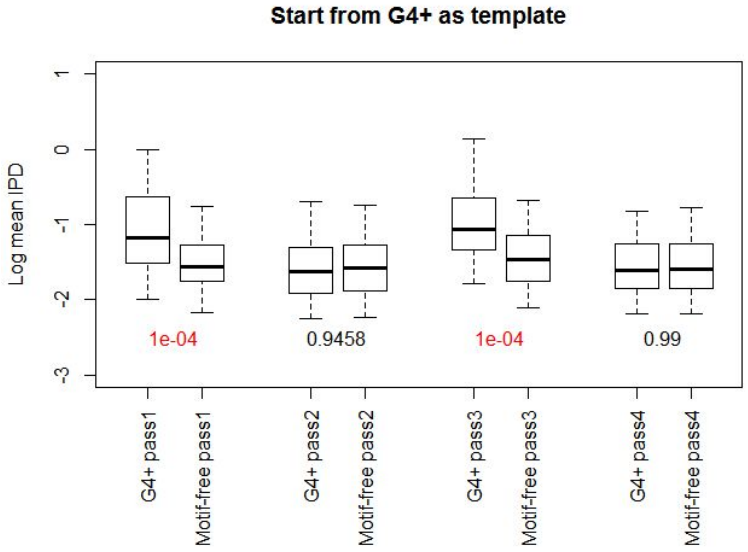

G−quadruplex − vs Motif−free
IPD reference strand
0
1
2
3
4
5
6
7
−40
−20
0
20
40
Motif center
Motif−free
median
25−75%
5−95%
G−quadruplex −
median
25−75%
5−95%
Depth ref
Motif
5463
Windows
0%
25%
50%
75%
100%
125%
Percentage

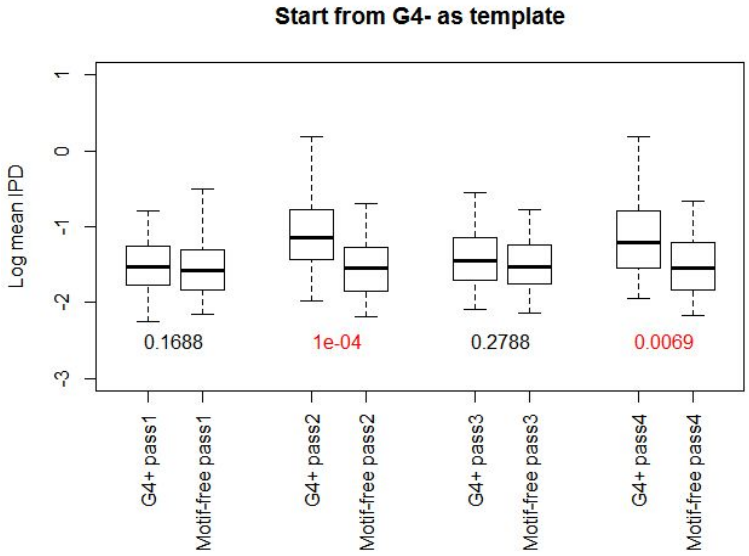
Figure S5. G-quadruplex structure is stable after multiple passes of sequencing of the circular template.

For every G4+ motif occurrence and matching motif-free region, we considered one molecule sequenced by exactly 4 passes (before polymerase drops, it uses G4+ as a template exactly twice), extracted the raw IPD information (using time between incorporation of consecutive bases in seconds) and computed the mean IPD. For each pass, we tested for differences between the mean IPD in G4+ and motif-free regions (two-sided test, multi-quantile statistic). We also tested for differences in mean IPDs between the first, and the second, the third, or the last (the 4th) pass in motif-free passes, finding no significance. **A** Molecules starting from G4+ as a template (142 molecules) versus motif-free passes. **B** Molecules starting from G4- as a template (115 molecules) versus motif-free passes. **C** Different motif-free passes. Boxplot whiskers mark the 5th and 95th quantiles. White: not significant (p-value>0.05). Red (Blue): significant with mean IPD higher (lower) in G4+ than motif-free regions. The analysis was performed on subsampled PacBio data with average depth of 12x.

A



B


Start from G4- as template
Log mean IPD
1
0
-1
-2
-3
0.1688
1e-04
0.2788
0.0069
G4+ pass1
Motif-free pass1
G4+ pass2
Motif-free pass2
G4+ pass3
Motif-free pass3
G4+ pass4
Motif-free pass4

C

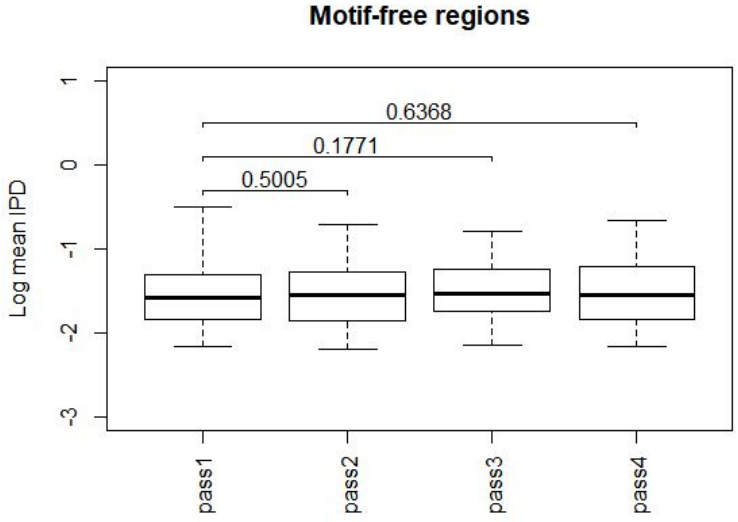
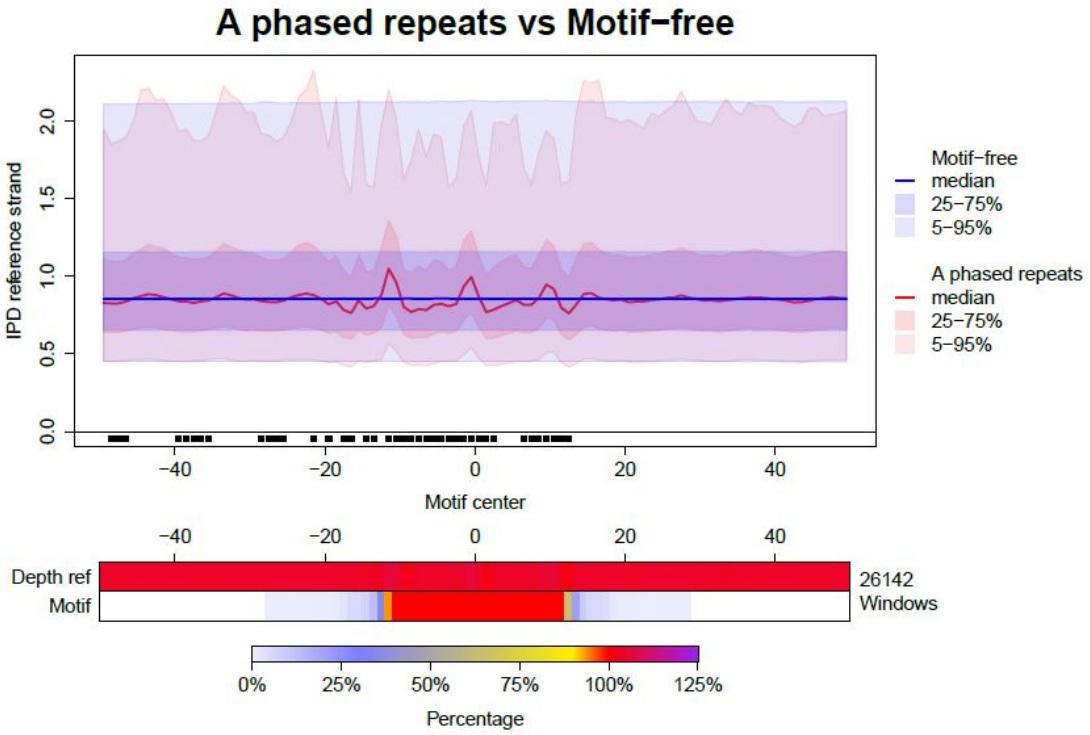

Motif-free regions
Log mean IPD
1
0
-1
-2
-3
0.6368
0.1771
0.5005
pass1
pass2
pass3
pass4

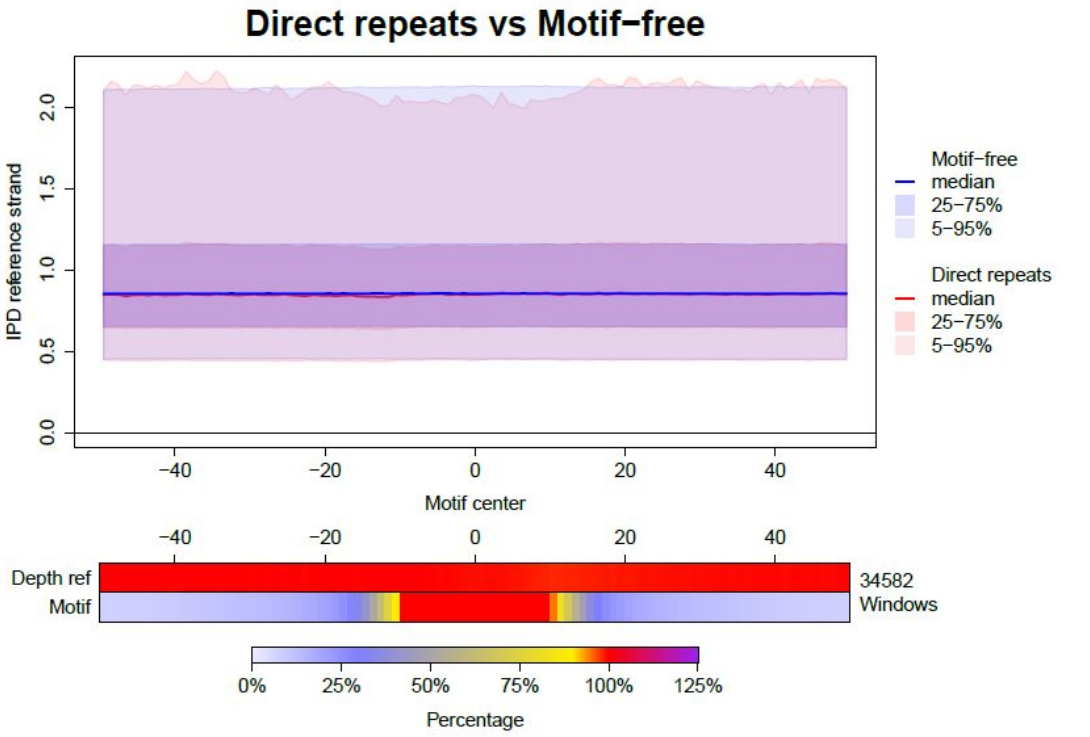
Figure S6. Effect of different non-B DNA motifs on IPDs.

A A-phased repeats depress the IPD distribution. **B** Direct Repeats do not significantly change the IPD distribution. **C** Inverted Repeats depress the IPD distribution slightly. **D** Mirror Repeats slightly depress the IPD distribution. **E** Z-DNA motifs slightly increase the IPD distribution in both strands. See the legend of Fig. 2A for details.

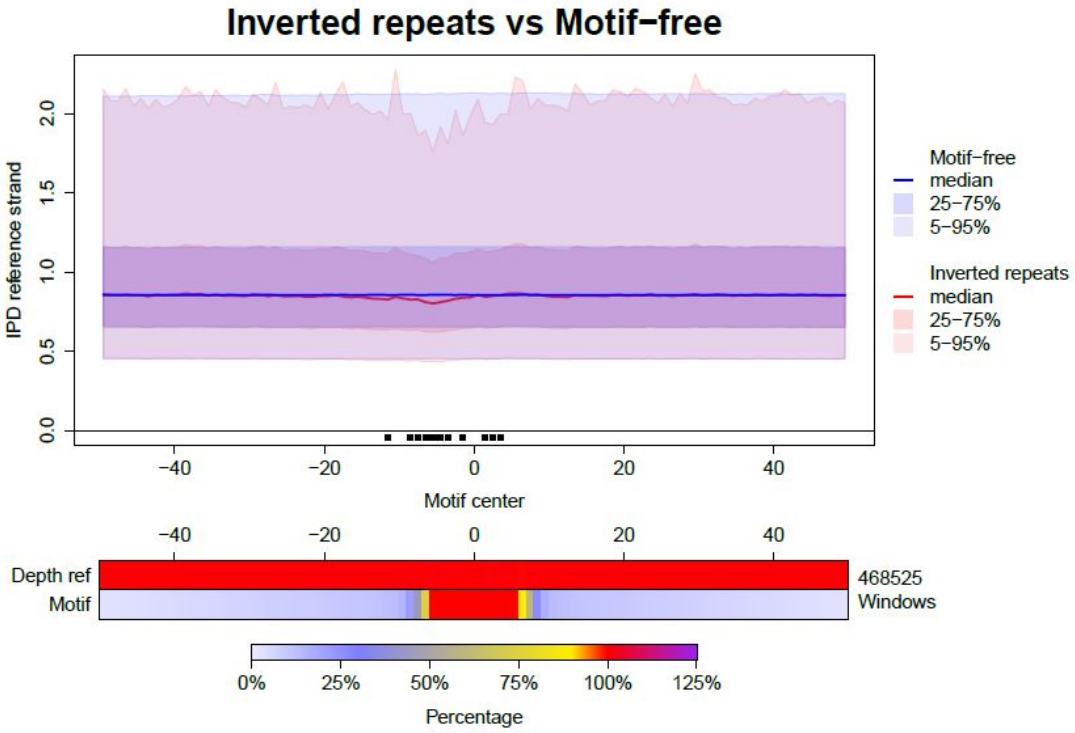
A



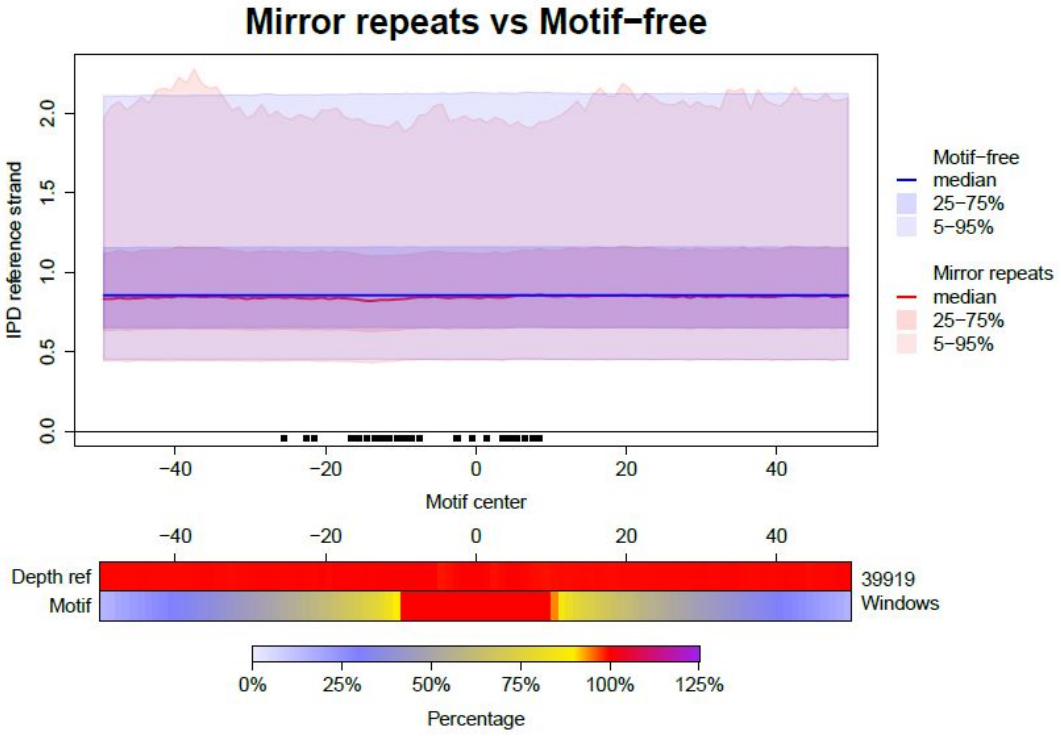
B



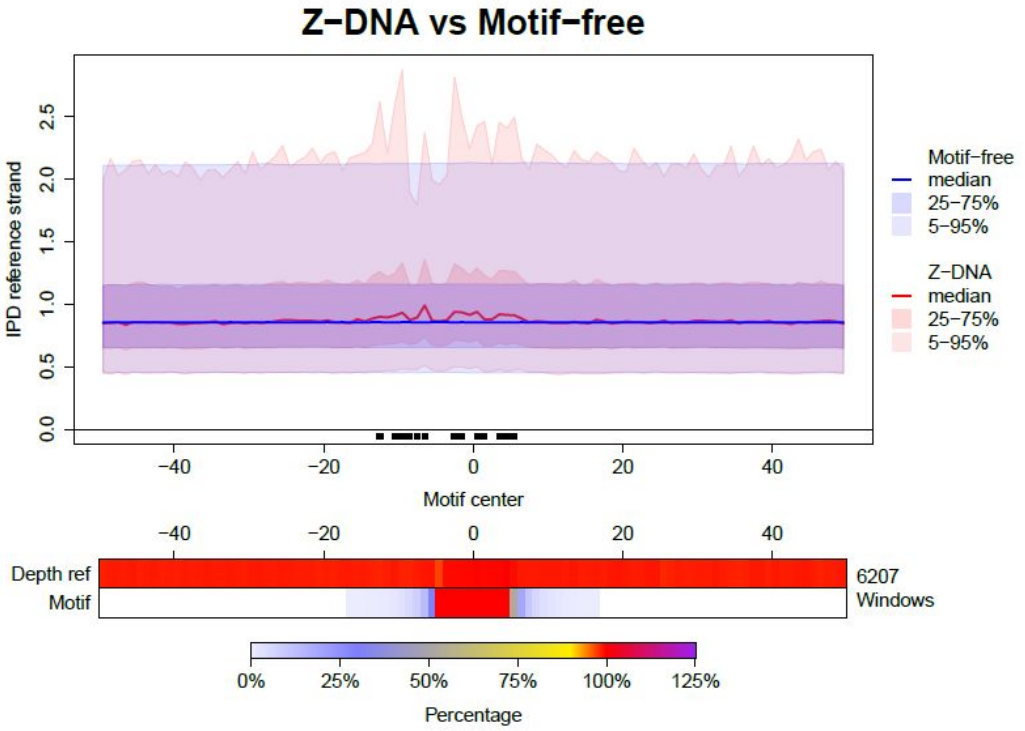
C


Inverted repeats vs Motif−free
IPD reference strand
0.0
0.5
1.0
1.5
2.0
−40
−20
0
20
40
Motif center
Depth ref
Motif
468525
Windows
0%
25%
50%
75%
100%
125%
Percentage
Motif−free
median
25−75%
5−95%
Inverted repeats
median
25−75%
5−95%

D


Mirror repeats vs Motif−free
IPD reference strand
0.0
0.5
1.0
1.5
2.0
−40
−20
0
20
40
Motif center
Depth ref
Motif
39919
Windows
0%
25%
50%
75%
100%
125%
Percentage
Motif−free
median
25−75%
5−95%
Mirror repeats
median
25−75%
5−95%

E


Z−DNA vs Motif−free
IPD reference strand
0.0
0.5
1.0
1.5
2.0
2.5
−40
−20
0
20
40
Motif center
Motif−free
median
25−75%
5−95%
Z−DNA
median
25−75%
5−95%
Depth ref
Motif
6207
Windows
0%
25%
50%
75%
100%
125%
Percentage

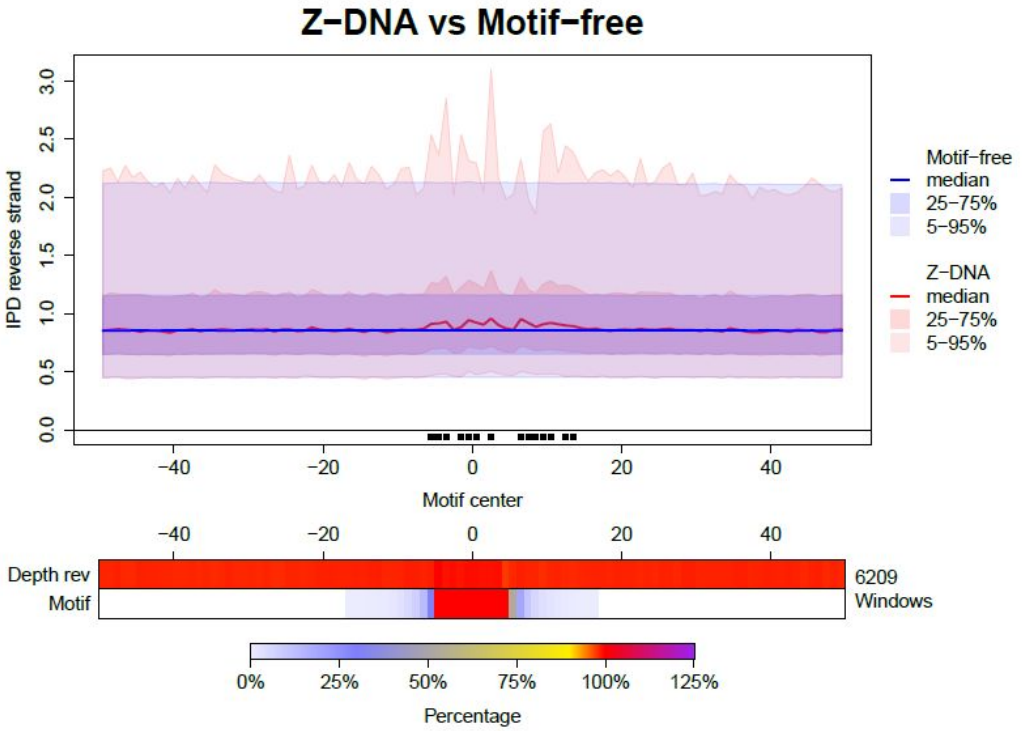
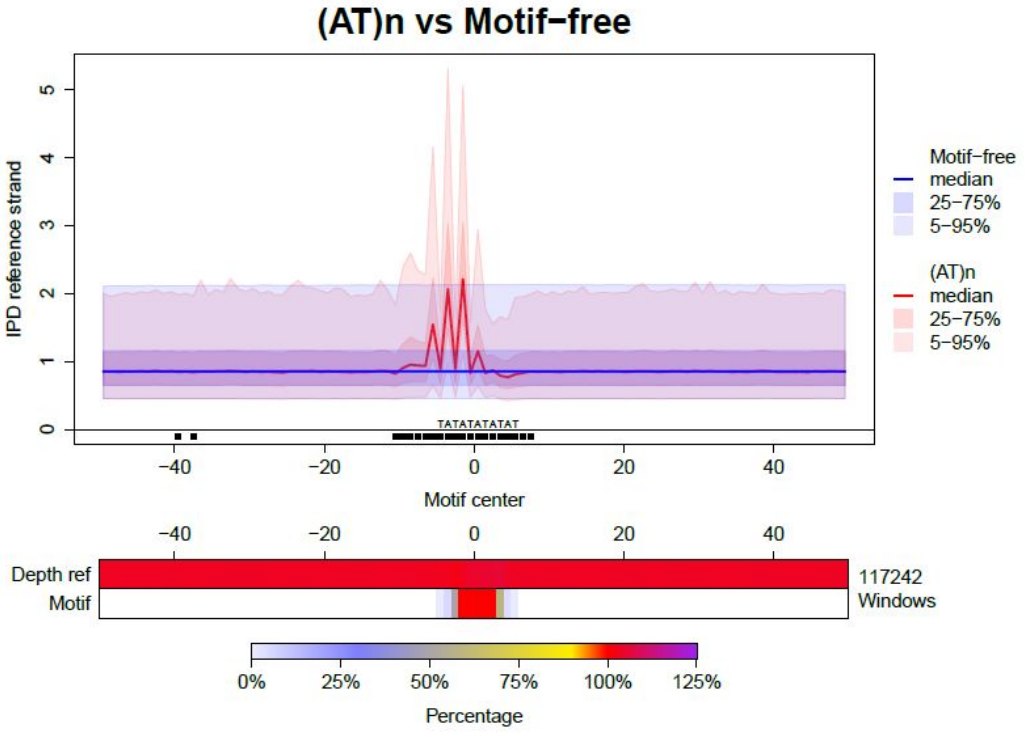
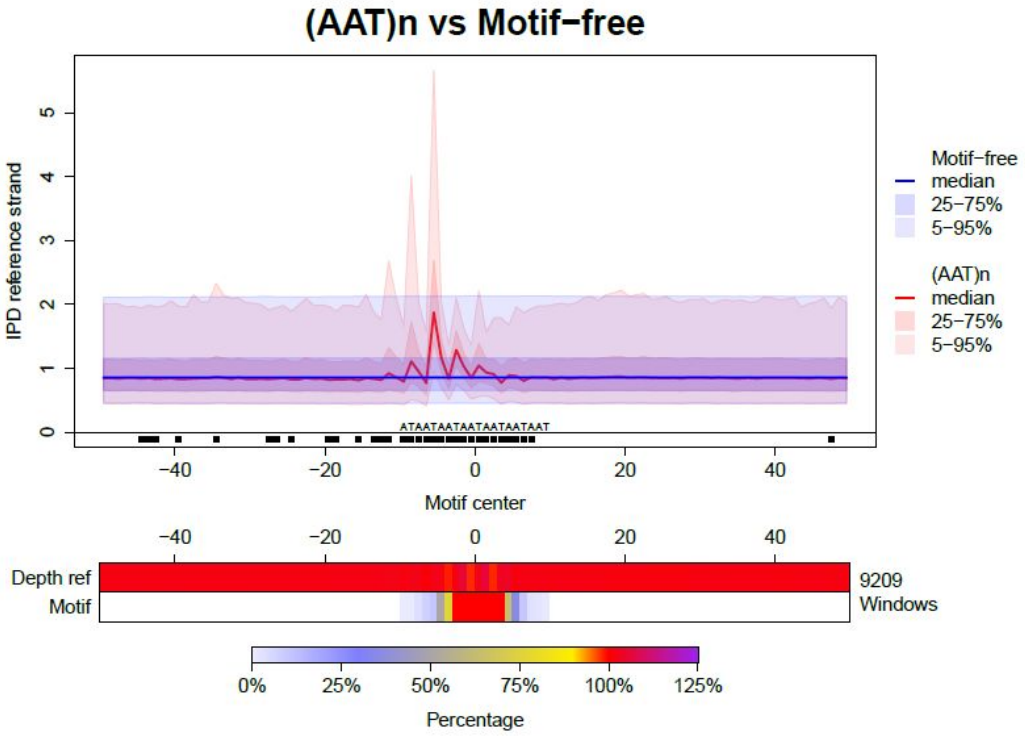

Z−DNA vs Motif−free
IPD reverse strand
0.0
0.5
1.0
1.5
2.0
2.5
3.0
−40
−20
0
20
40
Motif center
Motif−free
median
25−75%
5−95%
Z−DNA
median
25−75%
5−95%
Depth rev
Motif
6209
Windows
0%
25%
50%
75%
100%
125%
Percentage

Figure S7. The effect of STRs that can form hairpins on polymerization kinetics.
A $(AT)_n$. **B** $(AAT)_n$. **C** $(ACT)_n$. **D** $(AGG)_n$. **E** $(AGT)_n$. **F** $(ATC)_n$. **G** $(ATG)_n$. **H** $(ATT)_n$. See the legend of Fig. 2A.

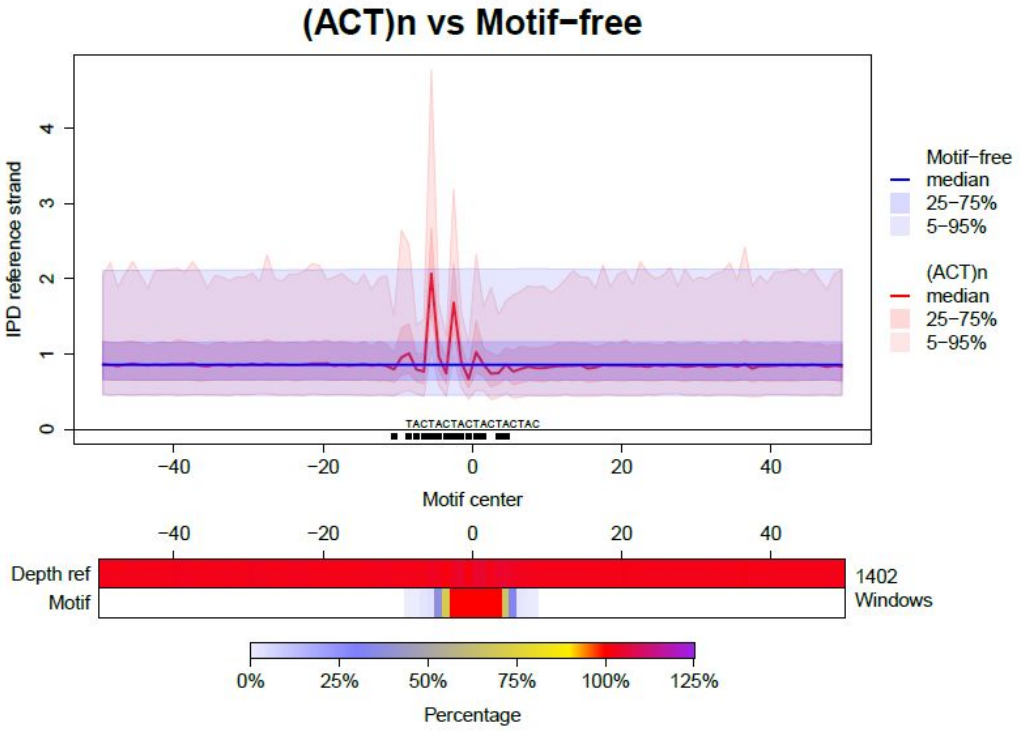
A



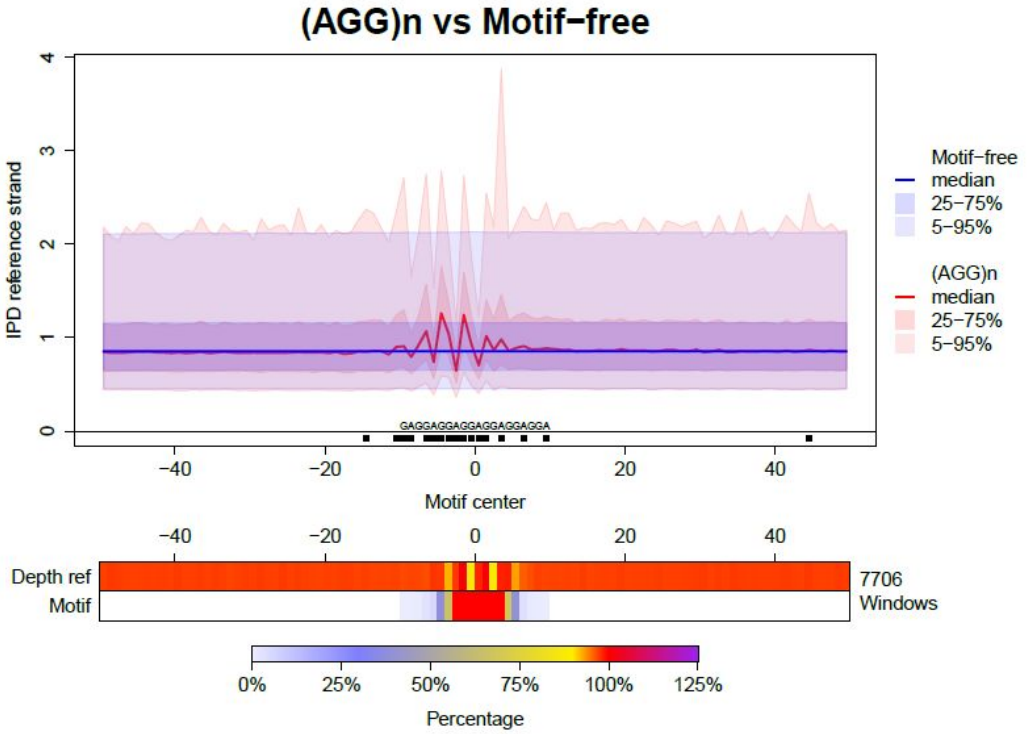
B



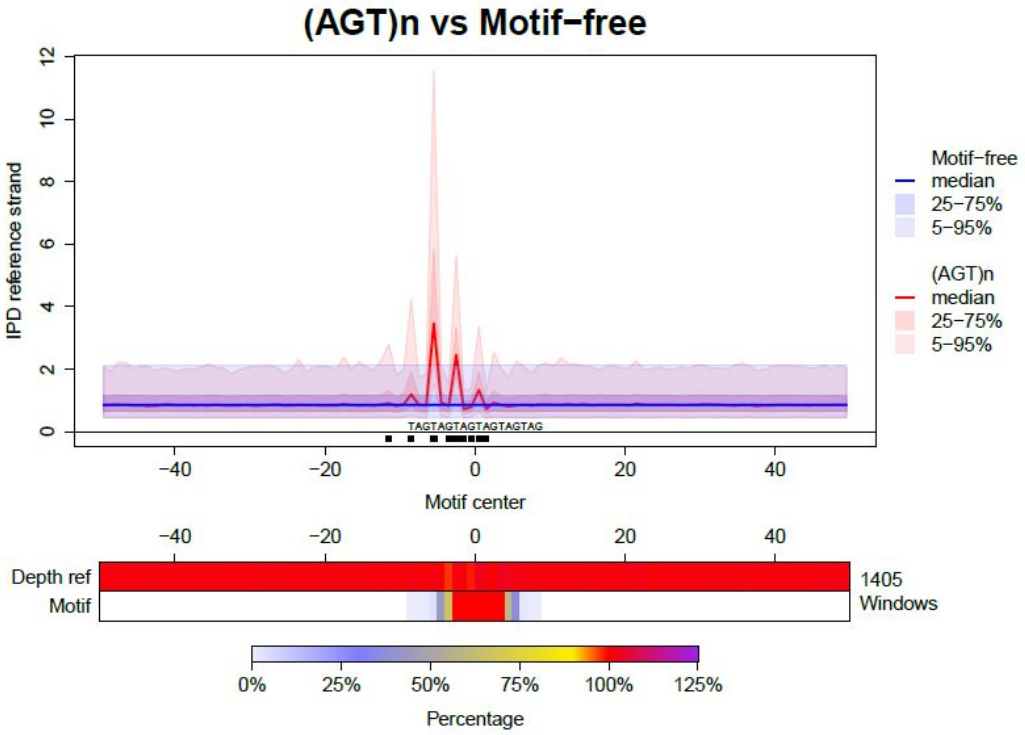
C


(ACT)n vs Motif−free
IPD reference strand
TACTACTACTACTAC
Motif center
Depth ref
Motif
1402
Windows
Motif−free
median
25−75%
5−95%
(ACT)n
median
25−75%
5−95%
0%
25%
50%
75%
100%
125%
Percentage

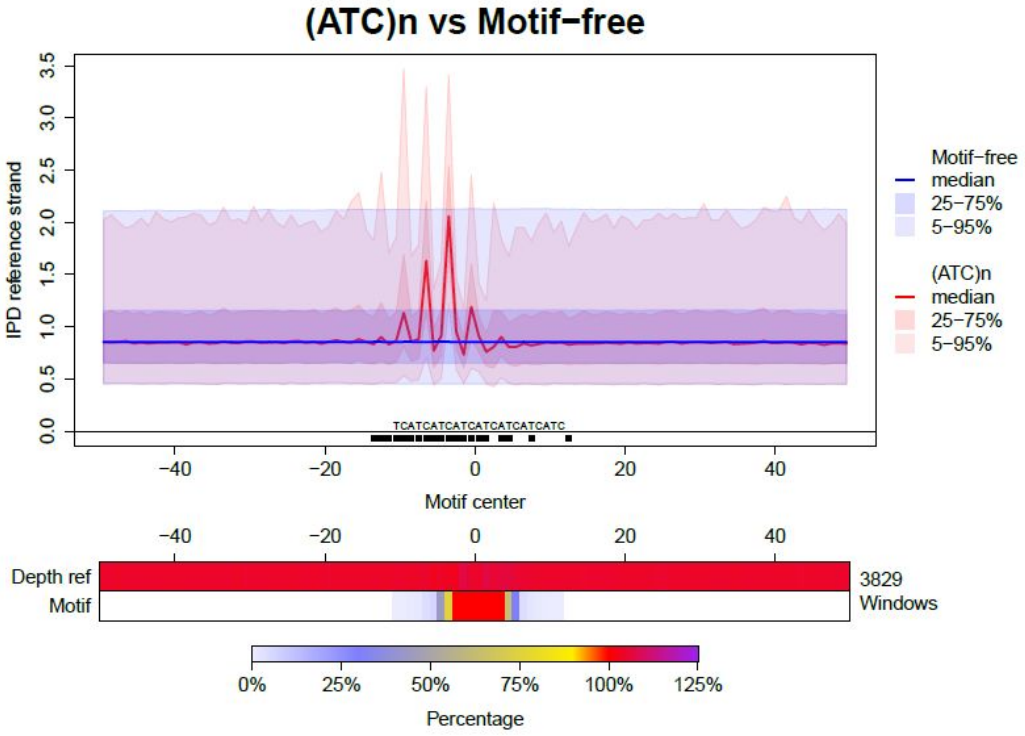
D


(AGG)n vs Motif−free
IPD reference strand
GAGGAGGAGGAGGAGGAGGA
Motif center
Depth ref
Motif
7706
Windows
Motif−free
median
25−75%
5−95%
(AGG)n
median
25−75%
5−95%
0%
25%
50%
75%
100%
125%
Percentage

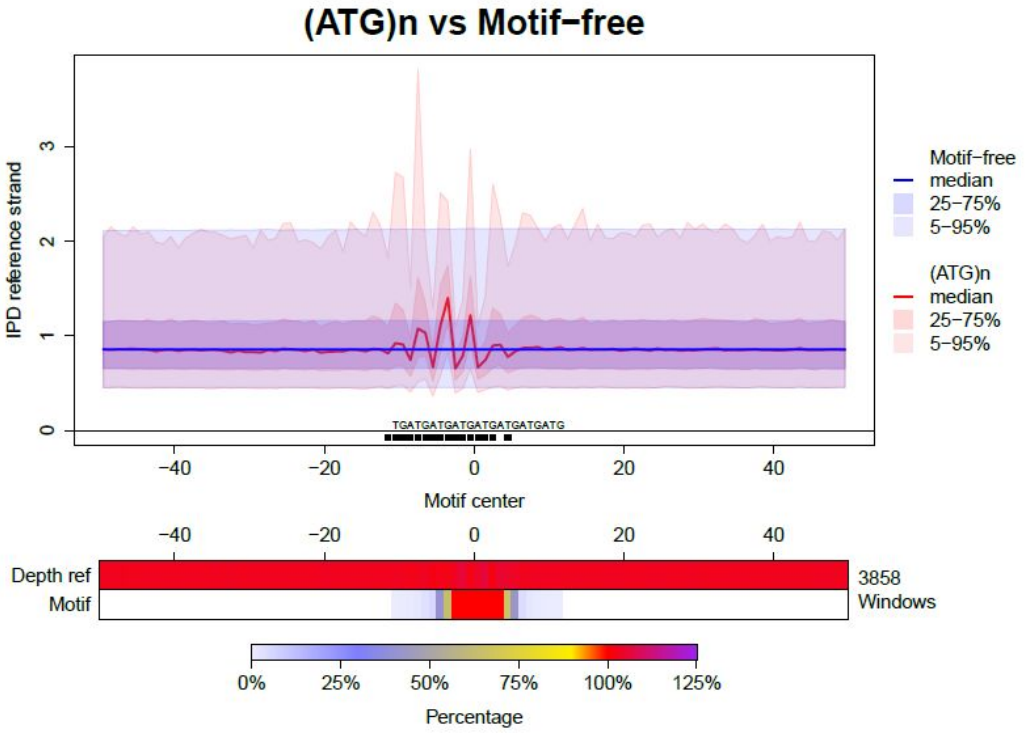
E


(AGT)n vs Motif−free
IPD reference strand
0
2
4
6
8
10
12
TAGTAGTAGTAGTAGTAG
−40
−20
0
20
40
Motif center
Depth ref
Motif
1405
Windows
0%
25%
50%
75%
100%
125%
Percentage
Motif−free
median
25−75%
5−95%
(AGT)n
median
25−75%
5−95%

F


(ATC)n vs Motif−free
IPD reference strand
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
TCATCATCATCATCATCATC
−40
−20
0
20
40
Motif center
Depth ref
Motif
3829
Windows
0%
25%
50%
75%
100%
125%
Percentage
Motif−free
median
25−75%
5−95%
(ATC)n
median
25−75%
5−95%

G


(ATG)n vs Motif−free
IPD reference strand
TGATGATGATGATGATGATG
Motif center
Depth ref
Motif
3858
Windows
Motif−free
median
25−75%
5−95%
(ATG)n
median
25−75%
5−95%
0%
25%
50%
75%
100%
125%
Percentage

H

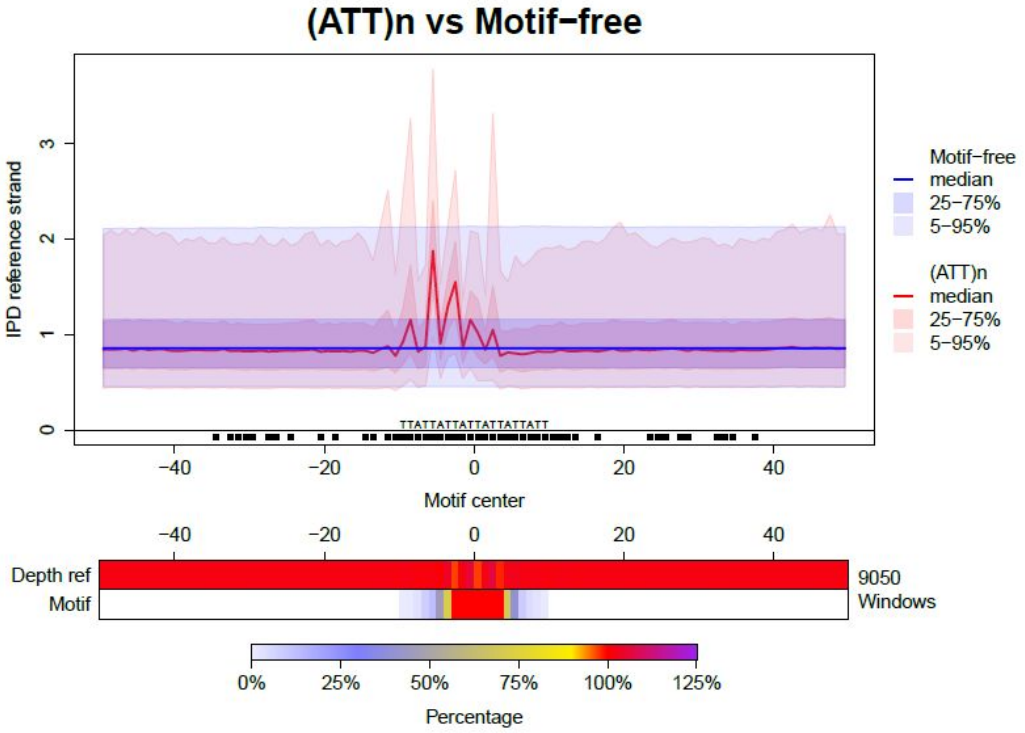
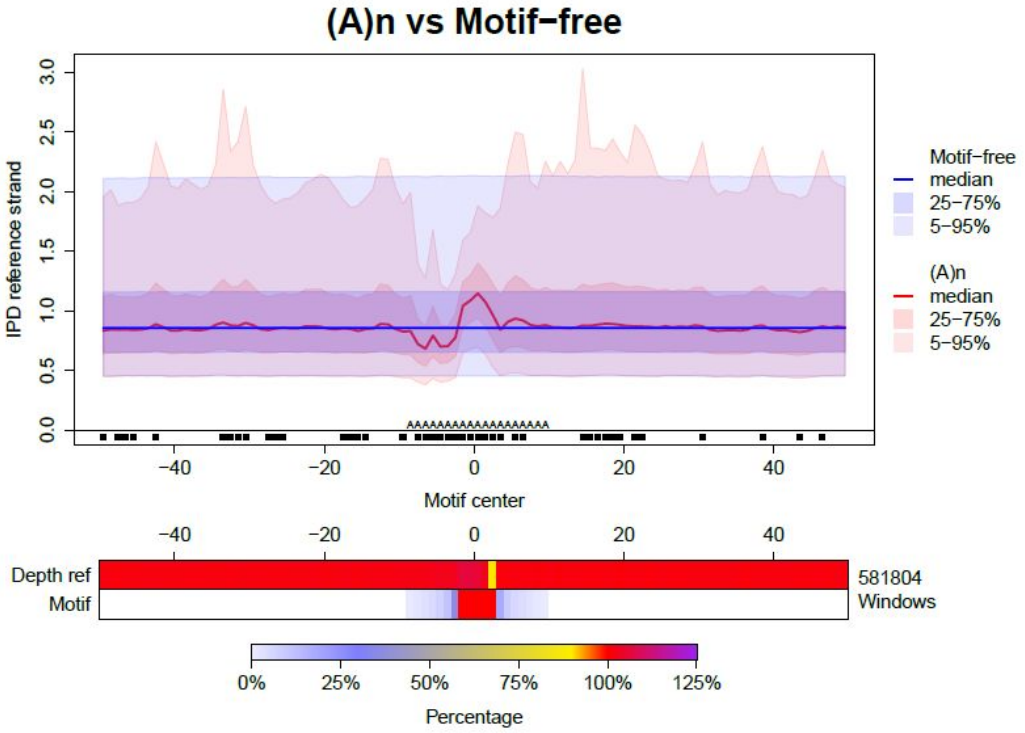

(ATT)n vs Motif−free
IPD reference strand
TTATTATTATTATTATTATT
Motif center
Depth ref
Motif
9050
Windows
Motif−free
median
25−75%
5−95%
(ATT)n
median
25−75%
5−95%
0%
25%
50%
75%
100%
125%
Percentage

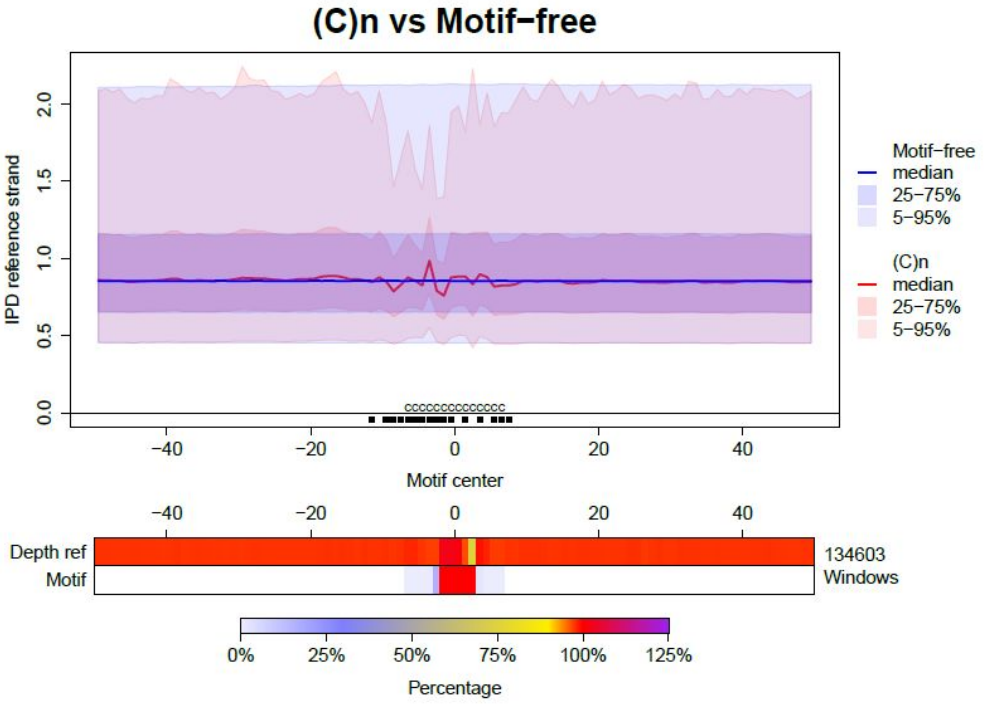
Figure S8. The effect of homopolymers and STRs that can form H-DNA on polymerization kinetics.

A $(A)_n$. **B** $(C)_n$. **C** $(G)_n$. **D** $(T)_n$. **E** $(A)_n$ with different lengths. **F** $(T)_n$ with different lengths. **G** $(AG)_n$. **H** $(CT)_n$. **I** $(CCT)_n$. See the legend of Fig. 2A for details about panels A-D and G-J. Panels E-F show the summary of the IWT results (see caption of Fig. 2E for details) for the comparisons of motif-containing vs. motif-free windows, with motif-containing windows grouped by the number of nucleotides in the motif (excluding lengths with fewer than 10 windows). We did not perform the analysis for $(C)_n$ and $(G)_n$ of different lengths because they are too short (their length ranges from 5 to 14 nucleotides, but only ~0.4% of them, 611 $(C)_n$ and 570 $(G)_n$, have length >7 nt). The relationship between mean IPD in the 100-bp windows (on a logarithmic scale) and motif length was also analyzed for all non-B DNA motifs using boxplots (results not shown).

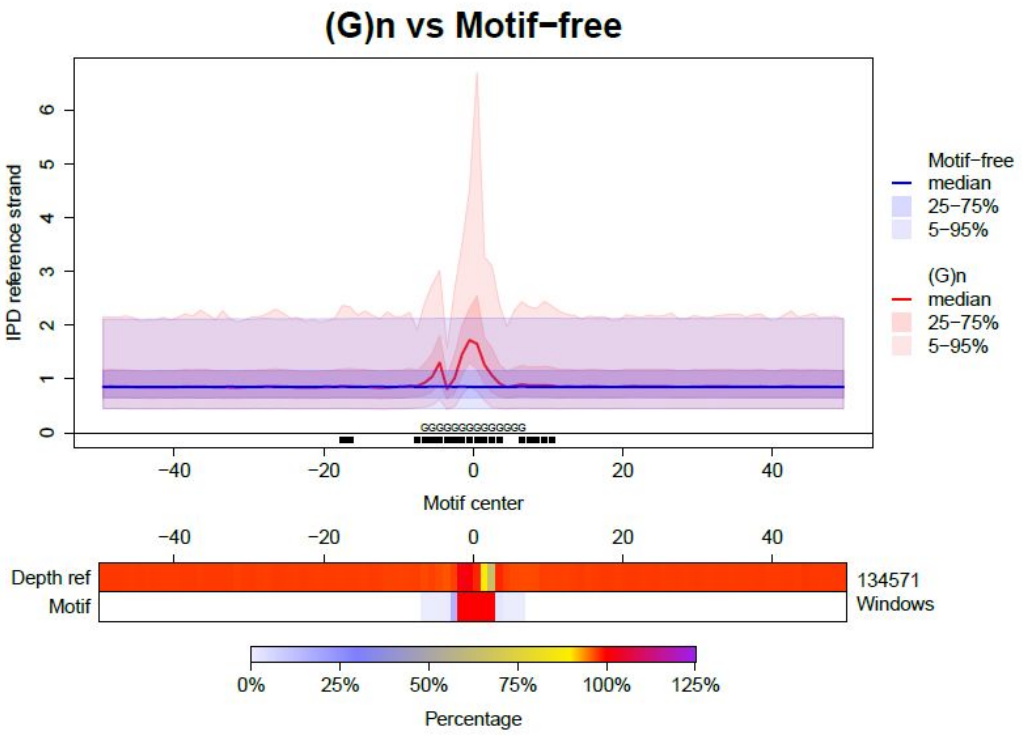
A



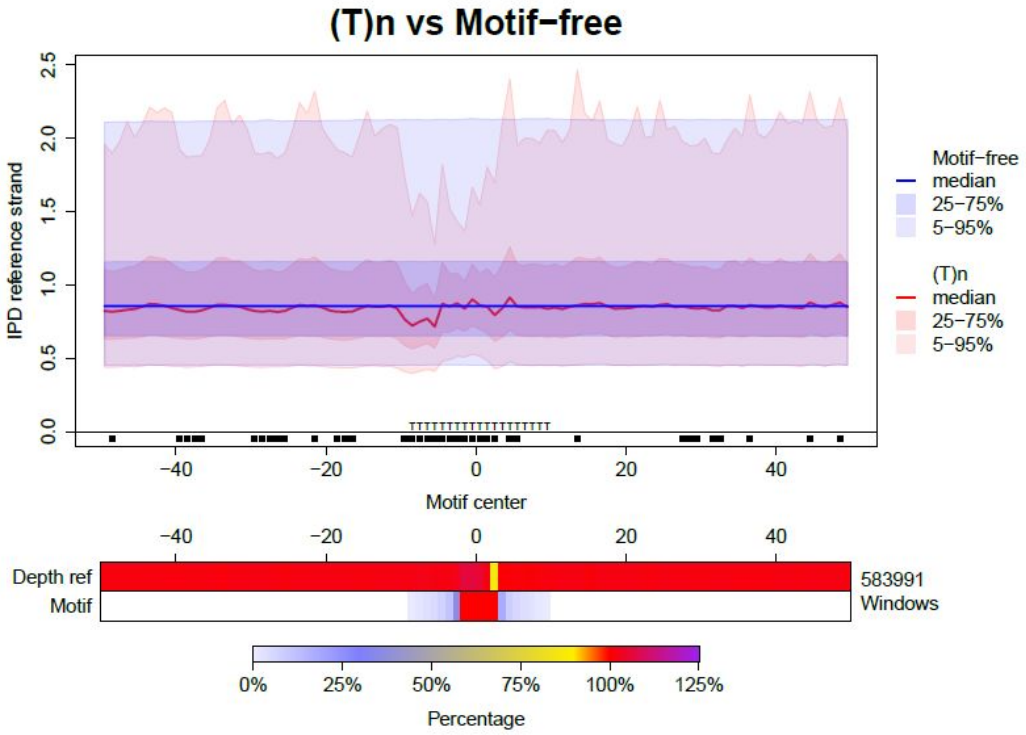
B


(C)n vs Motif−free
IPD reference strand
0.0
0.5
1.0
1.5
2.0
CCCCCCCCCCCCC
−40
−20
0
20
40
Motif center
Depth ref
Motif
134603
Windows
0%
25%
50%
75%
100%
125%
Percentage
Motif−free
median
25−75%
5−95%
(C)n
median
25−75%
5−95%

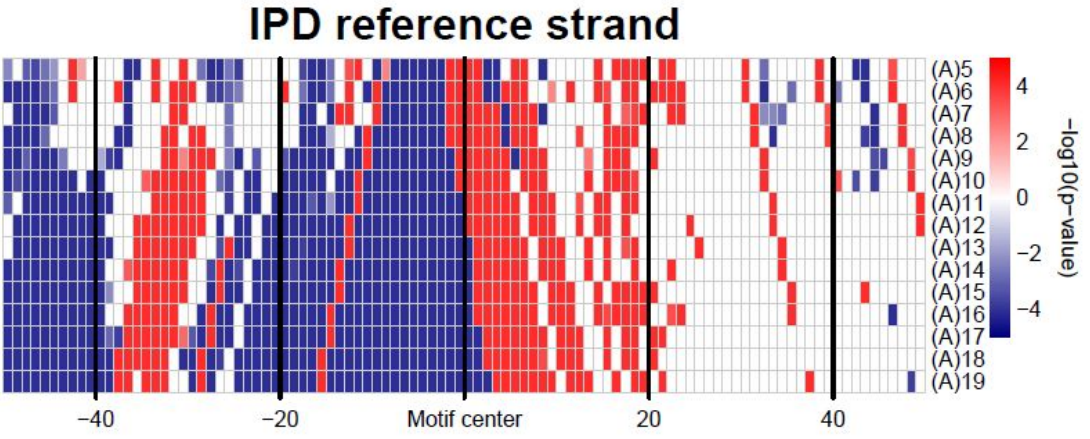
C


(G)n vs Motif−free
IPD reference strand
0
1
2
3
4
5
6
GGGGGGGGGGGGG
−40
−20
0
20
40
Motif center
Depth ref
Motif
134571
Windows
0%
25%
50%
75%
100%
125%
Percentage
Motif−free
median
25−75%
5−95%
(G)n
median
25−75%
5−95%

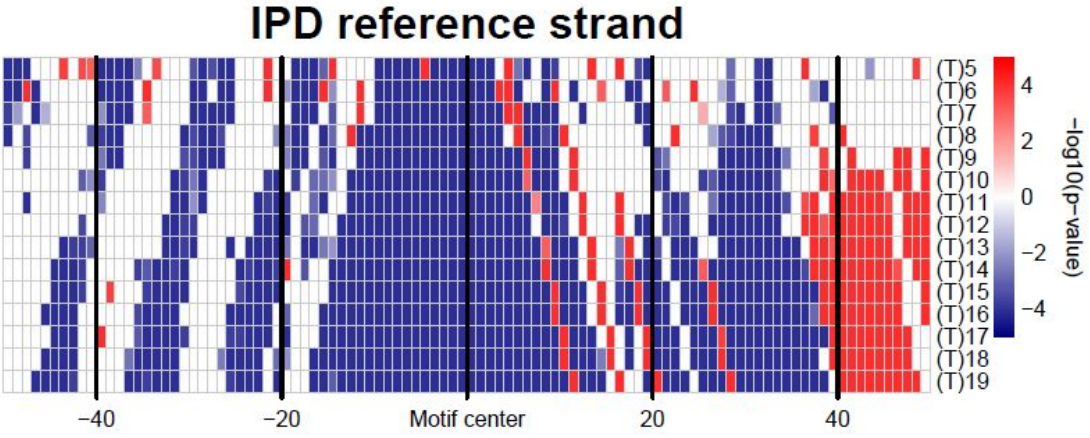
D


(T)n vs Motif−free
IPD reference strand
Motif−free
median
25−75%
5−95%
(T)n
median
25−75%
5−95%
Motif center
Depth ref
Motif
583991
Windows
0%
25%
50%
75%
100%
125%
Percentage

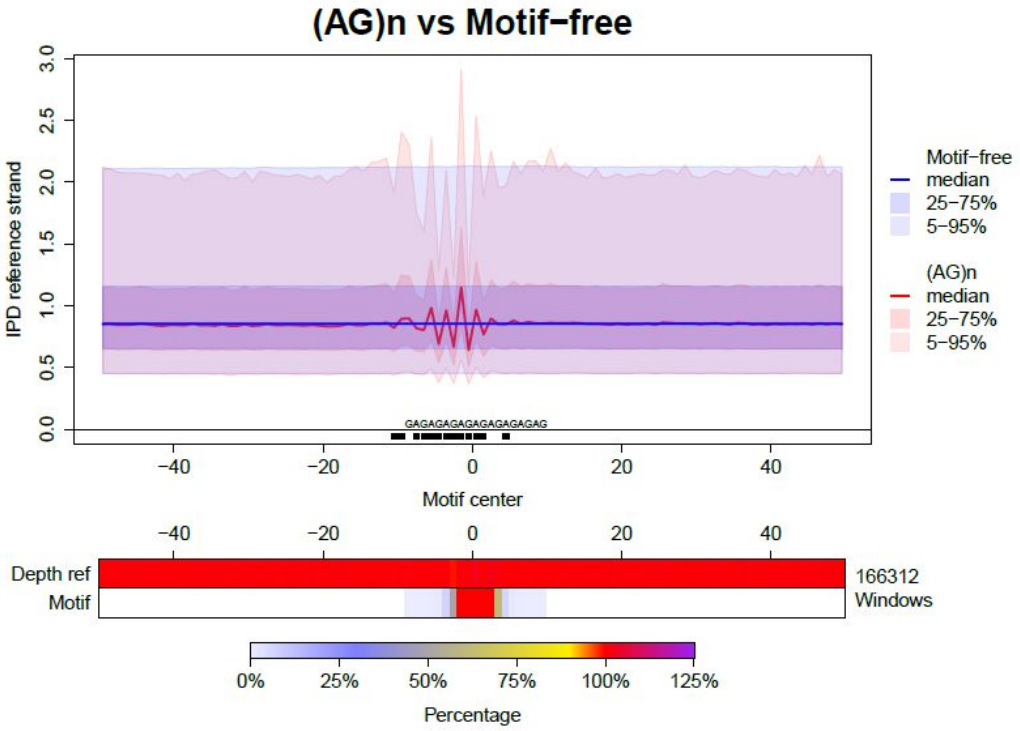
E


IPD reference strand
(A)5
(A)6
(A)7
(A)8
(A)9
(A)10
(A)11
(A)12
(A)13
(A)14
(A)15
(A)16
(A)17
(A)18
(A)19
−log10(p−value)
Motif center

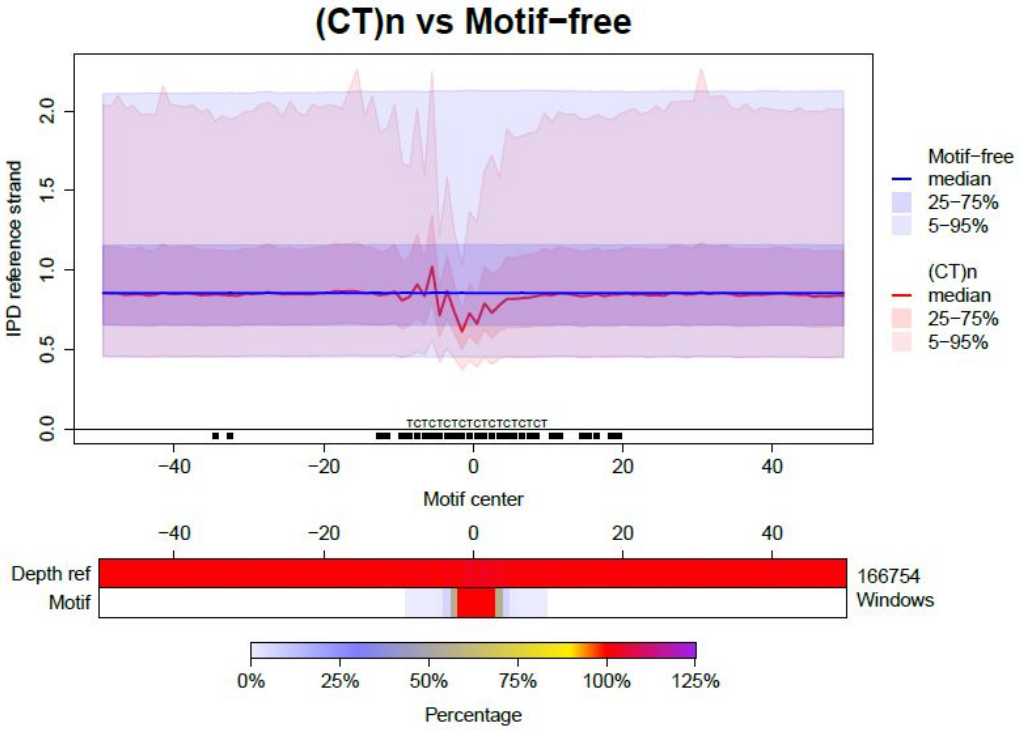
F


IPD reference strand
(T)5
(T)6
(T)7
(T)8
(T)9
(T)10
(T)11
(T)12
(T)13
(T)14
(T)15
(T)16
(T)17
(T)18
(T)19
−log10(p−value)
Motif center

G


(AG)n vs Motif−free
IPD reference strand
0.0
0.5
1.0
1.5
2.0
2.5
3.0
GAGAGAGAGAGAGAGAG
−40
−20
0
20
40
Motif center
Depth ref
Motif
166312
Windows
0%
25%
50%
75%
100%
125%
Percentage
Motif−free
median
25−75%
5−95%
(AG)n
median
25−75%
5−95%

H


(CT)n vs Motif−free
IPD reference strand
0.0
0.5
1.0
1.5
2.0
TCTCTCTCTCTCTCTCT
−40
−20
0
20
40
Motif center
Depth ref
Motif
166754
Windows
0%
25%
50%
75%
100%
125%
Percentage
Motif−free
median
25−75%
5−95%
(CT)n
median
25−75%
5−95%

I

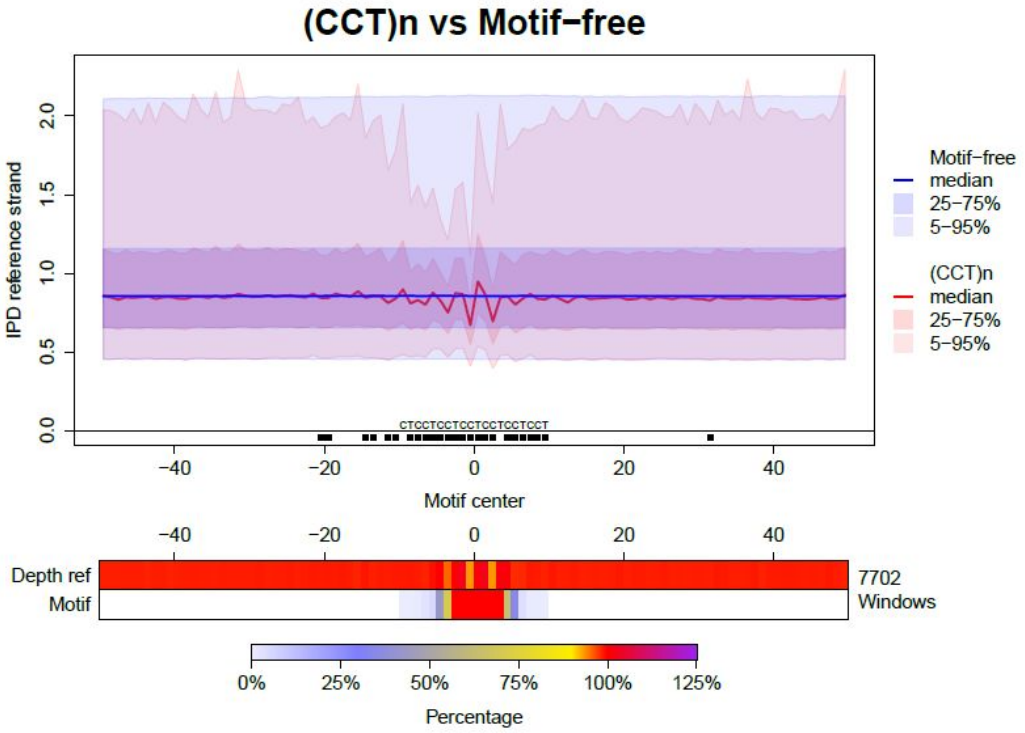
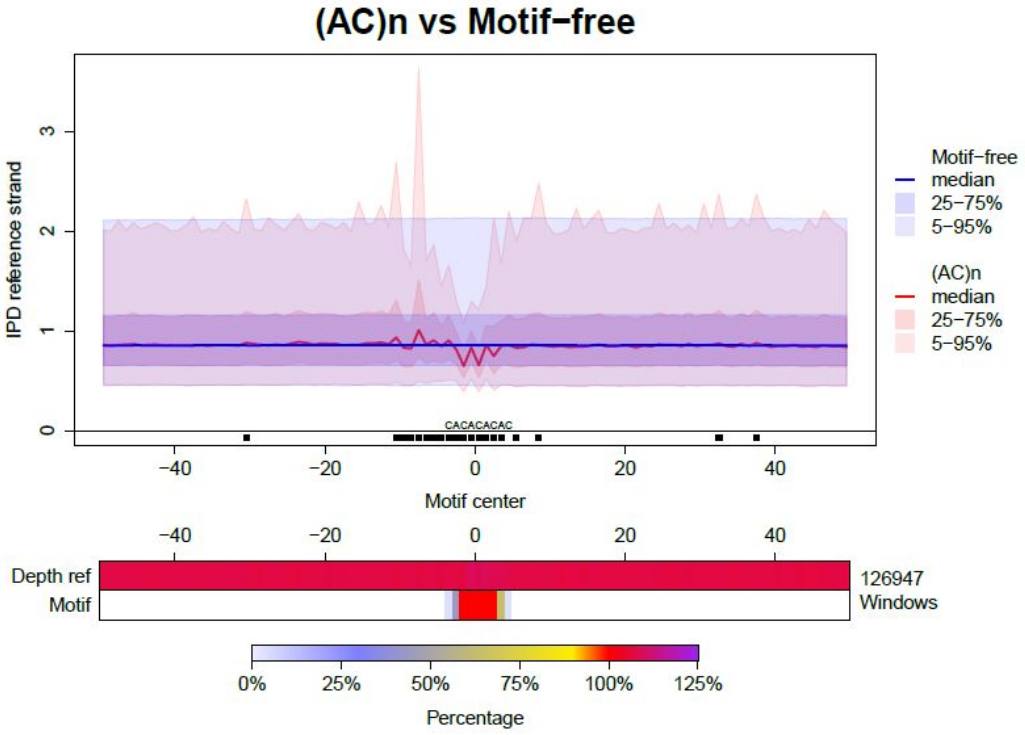
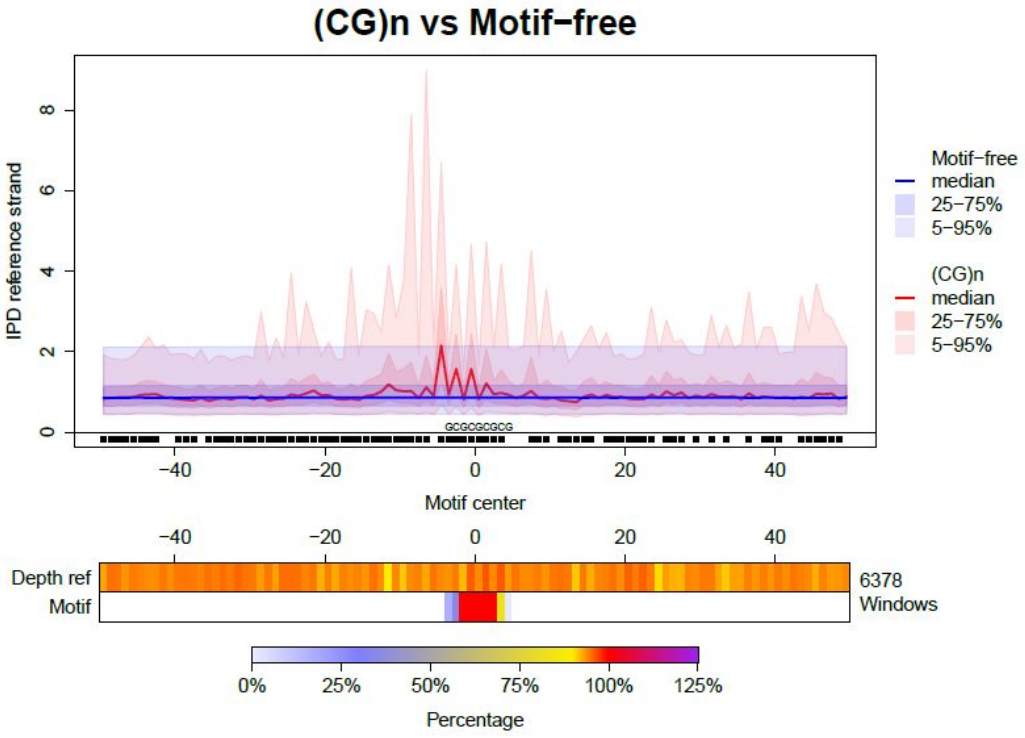

(CCT)n vs Motif−free
IPD reference strand
0.0
0.5
1.0
1.5
2.0
CTCCTCCTCCTCCTCCTCCT
−40
−20
0
20
40
Motif center
Motif−free
median
25−75%
5−95%
(CCT)n
median
25−75%
5−95%
Depth ref
Motif
7702
Windows
0%
25%
50%
75%
100%
125%
Percentage

Figure S9. The effect of STRs that can form Z-DNA on polymerization kinetics.
A $(AC)_n$. **B** $(CG)_n$. **C** $(GT)_n$. See the legend of Fig. 2A.

A



B



C

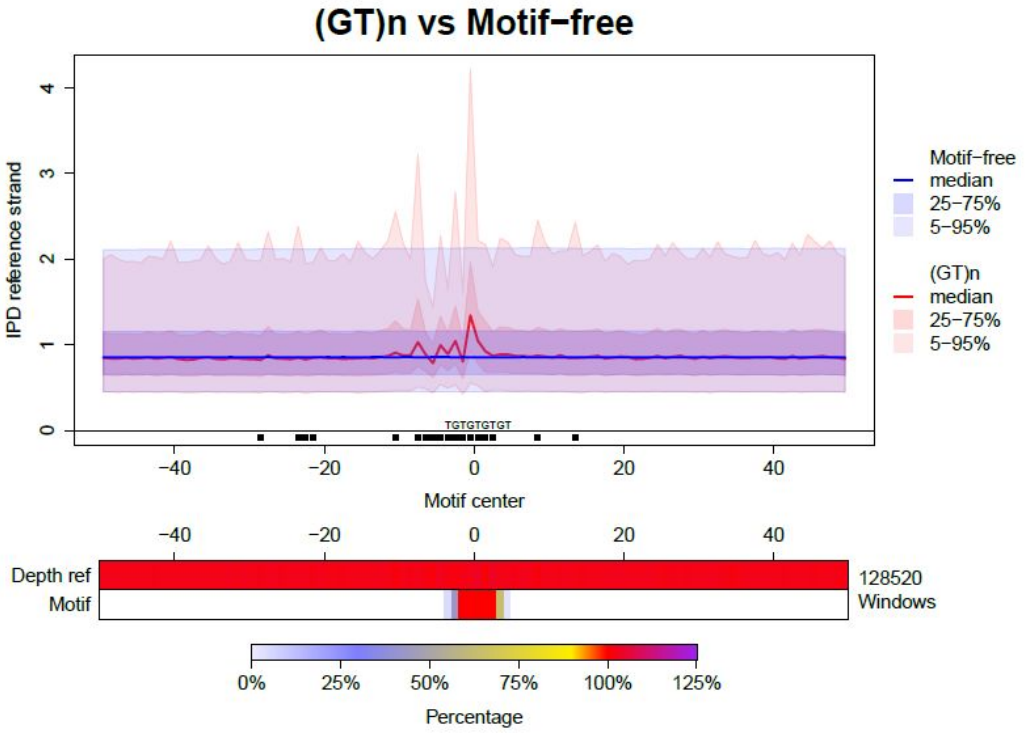
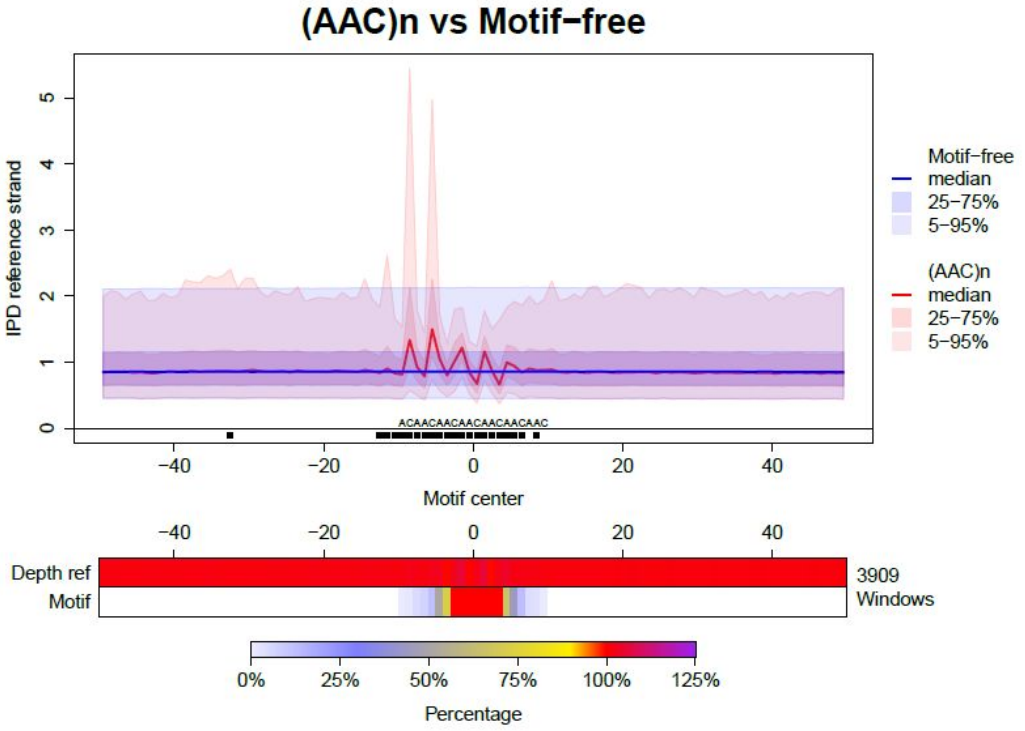
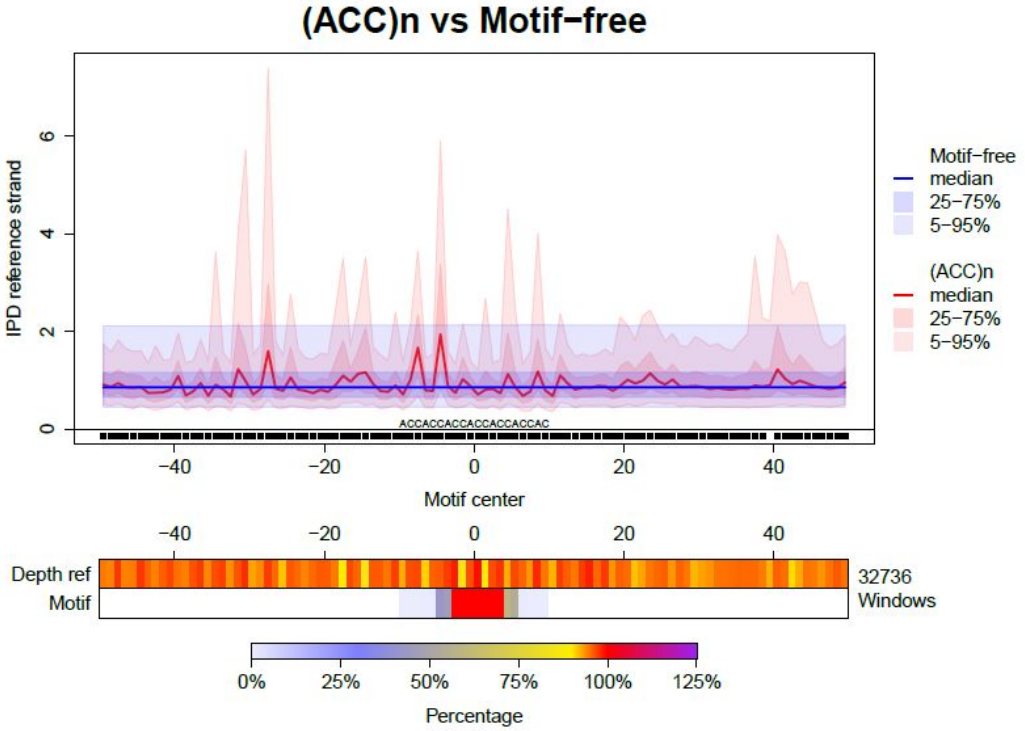

(GT)n vs Motif−free
IPD reference strand
0
1
2
3
4
TGTGTGTGT
−40
−20
0
20
40
Motif center
Motif−free
median
25−75%
5−95%
(GT)n
median
25−75%
5−95%
Depth ref
Motif
128520
Windows
0%
25%
50%
75%
100%
125%
Percentage

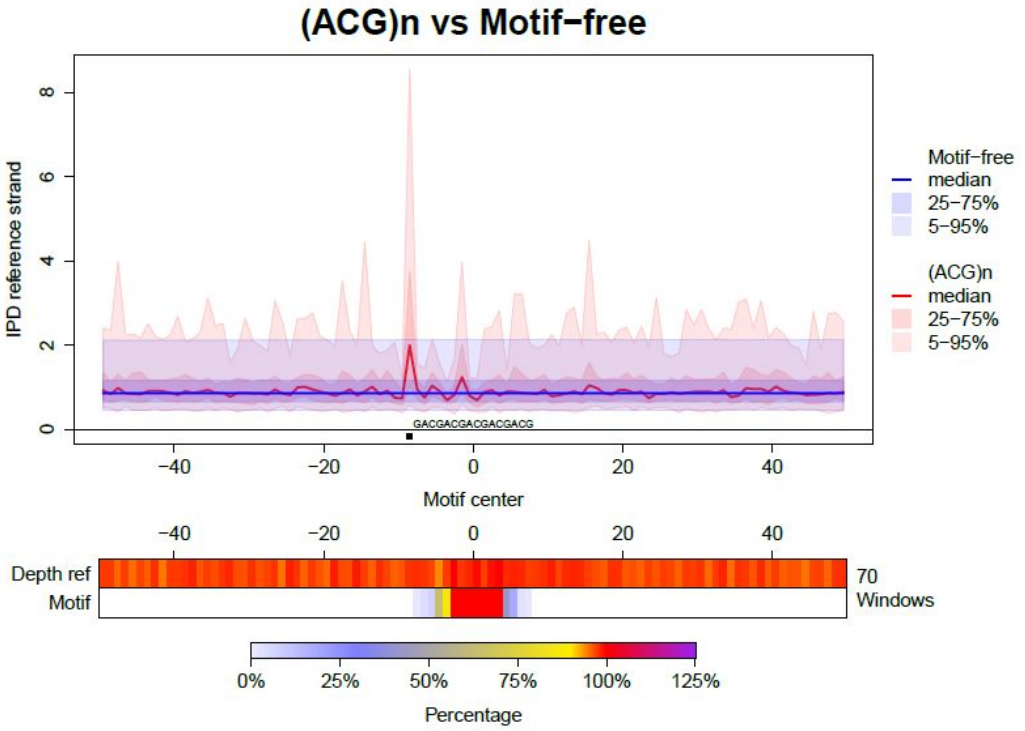
Figure S10. The effect of STRs on polymerization kinetics.
A $(AAC)_n$. **B** $(ACC)_n$. **C** $(ACG)_n$. **D** $(CGT)_n$. **E** $(GGT)_n$. **F** $(GTT)_n$. See the legend of Fig. 2A.
A



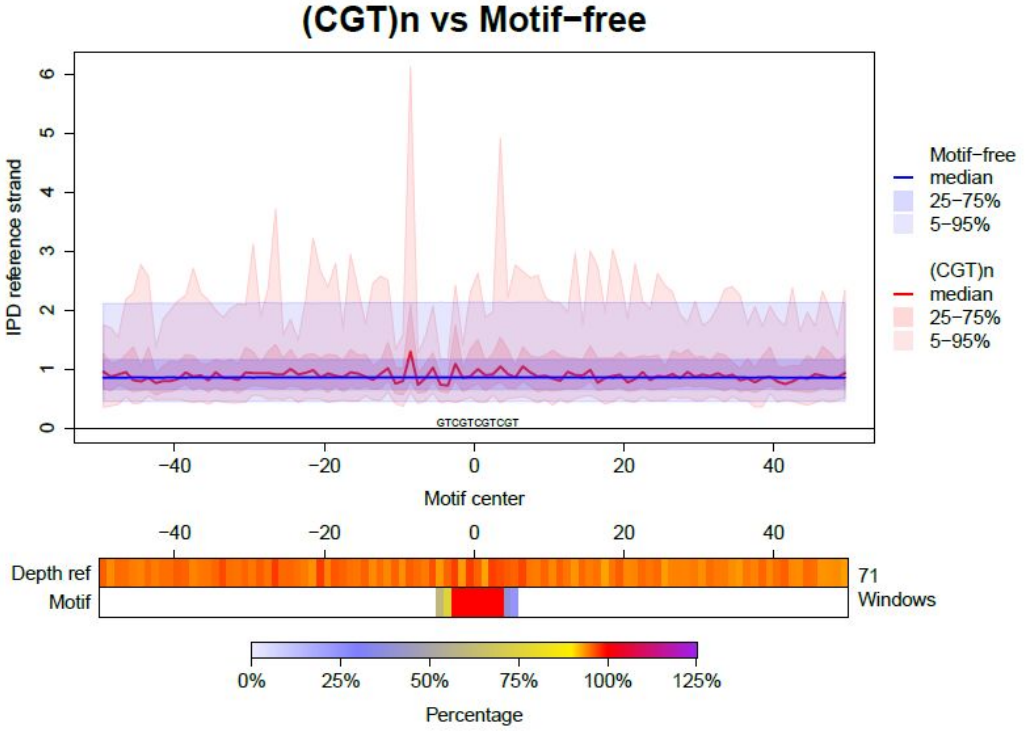
B



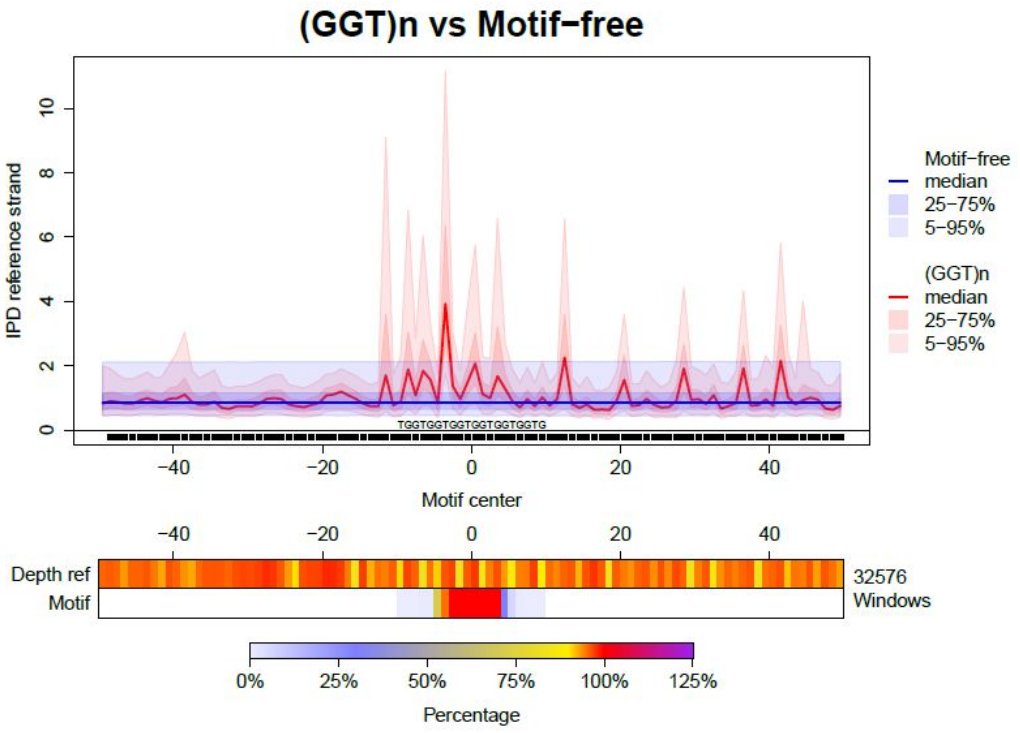
C


(ACG)n vs Motif−free
IPD reference strand
0
2
4
6
8
GACGACGACGACGACG
−40
−20
0
20
40
Motif center
Depth ref
Motif
70
Windows
0%
25%
50%
75%
100%
125%
Percentage
Motif−free
median
25−75%
5−95%
(ACG)n
median
25−75%
5−95%

D


(CGT)n vs Motif−free
IPD reference strand
0
1
2
3
4
5
6
GTCGTCGTCGT
−40
−20
0
20
40
Motif center
Depth ref
Motif
71
Windows
0%
25%
50%
75%
100%
125%
Percentage
Motif−free
median
25−75%
5−95%
(CGT)n
median
25−75%
5−95%

E


(GGT)n vs Motif−free
IPD reference strand
0
2
4
6
8
10
TGGTGGTGGTGGTGGTGG
−40
−20
0
20
40
Motif center
Motif−free
median
25−75%
5−95%
(GGT)n
median
25−75%
5−95%
Depth ref
Motif
32576
Windows
0%
25%
50%
75%
100%
125%
Percentage

F

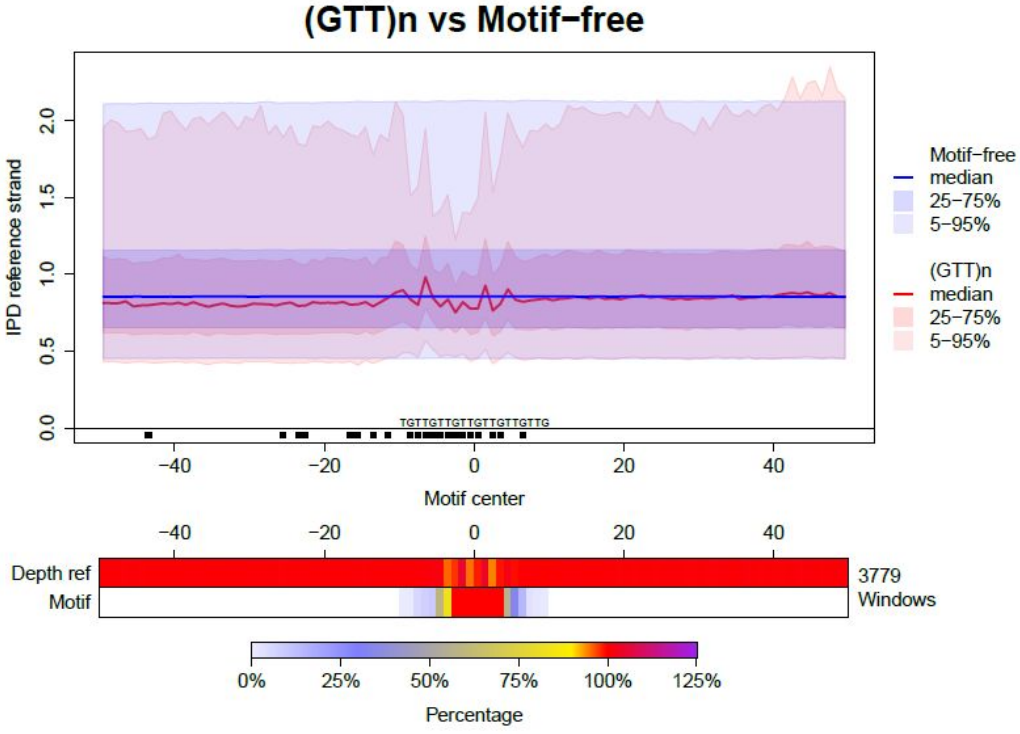
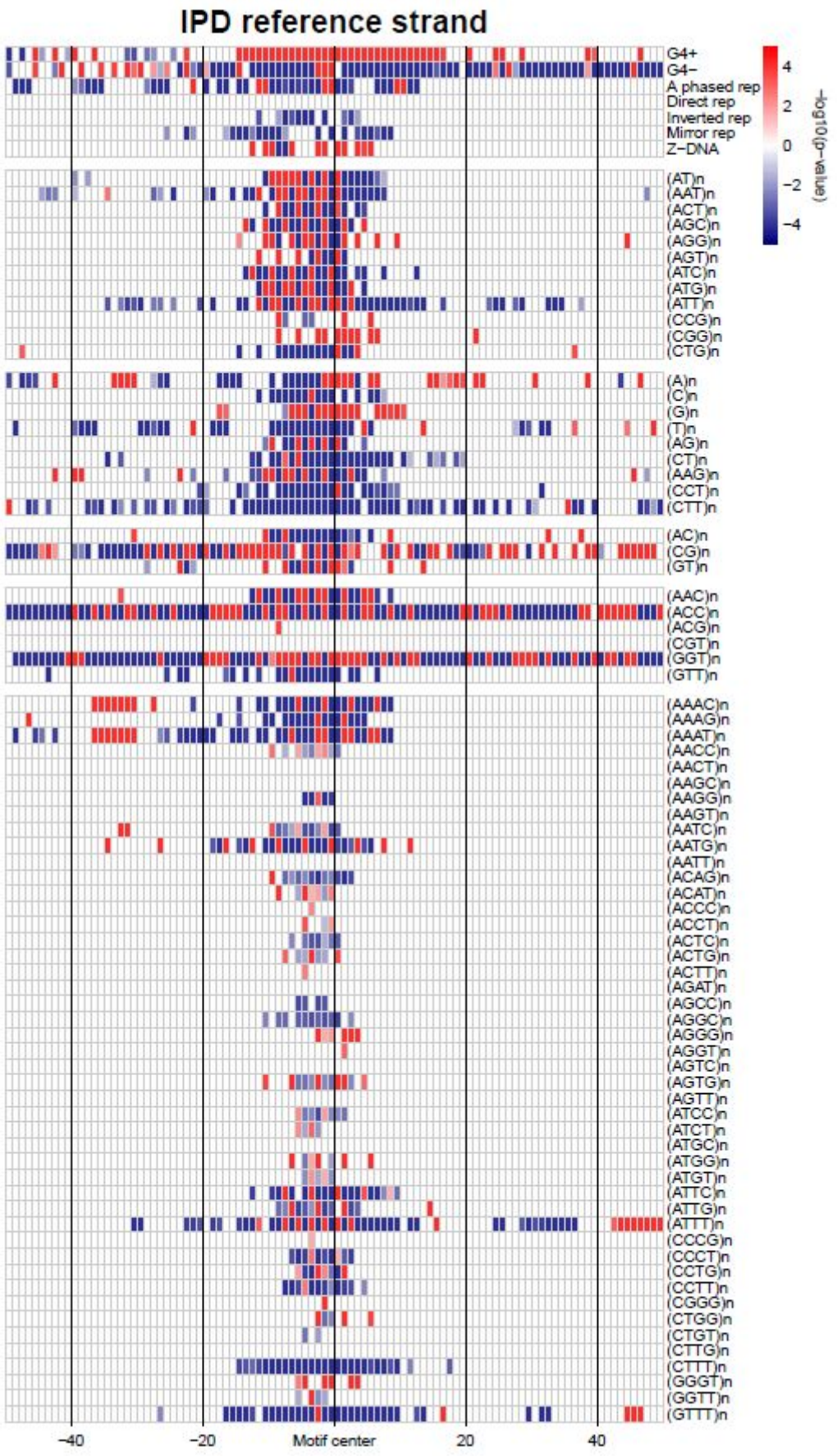
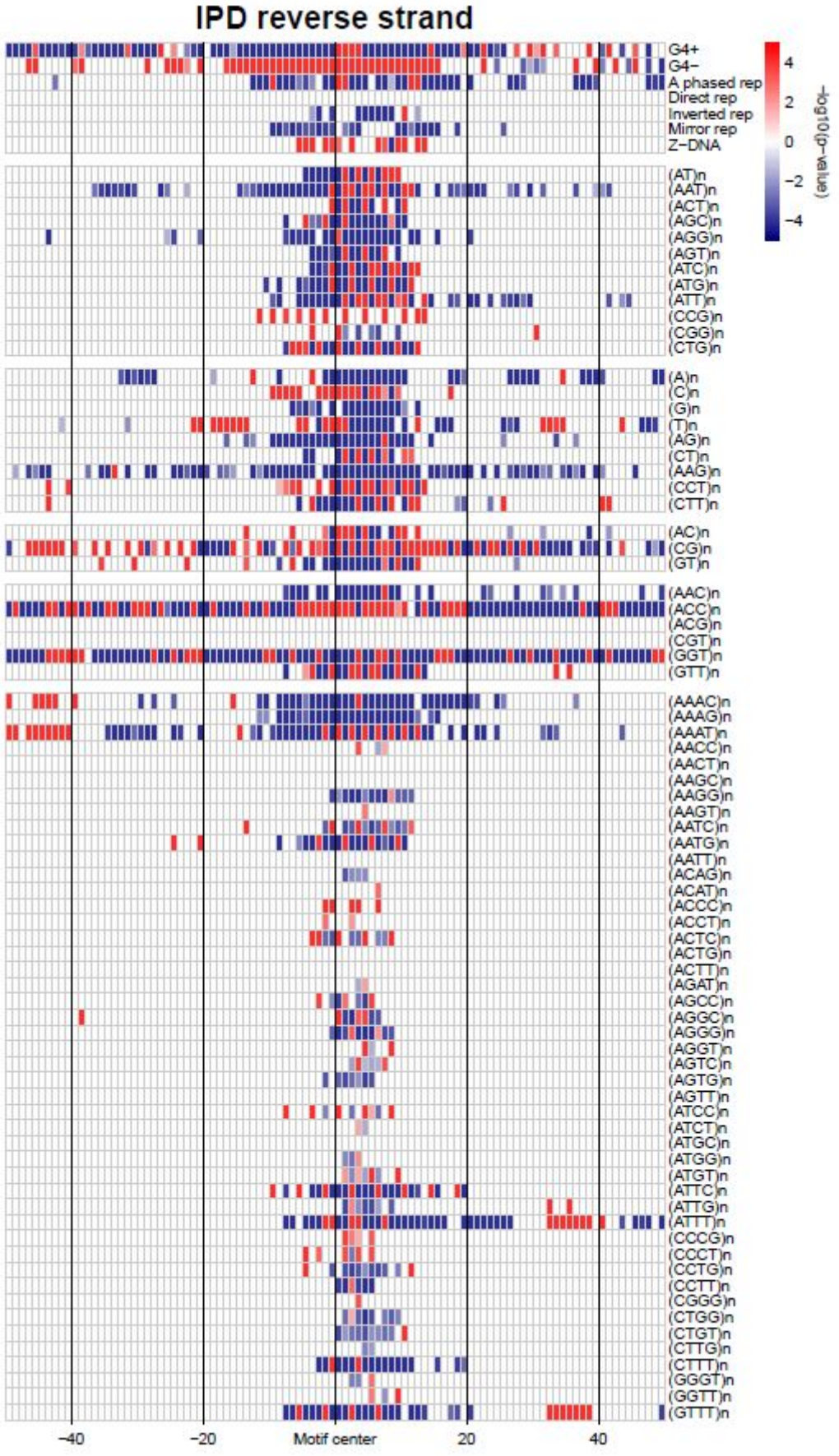

(GTT)n vs Motif−free
IPD reference strand
0.0
0.5
1.0
1.5
2.0
TGTTGTTGTTGTTGTTG
−40
−20
0
20
40
Motif center
Motif−free
median
25−75%
5−95%
(GTT)n
median
25−75%
5−95%
Depth ref
Motif
3779
Windows
0%
25%
50%
75%
100%
125%
Percentage

Figure S11. Summary of Interval-Wise Testing results for differences in IPDs.
A Reference strand, multi-quantile statistic. **B** Reverse complement strand, multi-quantile statistic. **C** Reference strand, mean statistic. **D** Reverse complement strand, mean statistic. **E** Reference strand, median statistic. **F** Reverse complement strand, median statistic. See the legend of Fig. 2E for details.

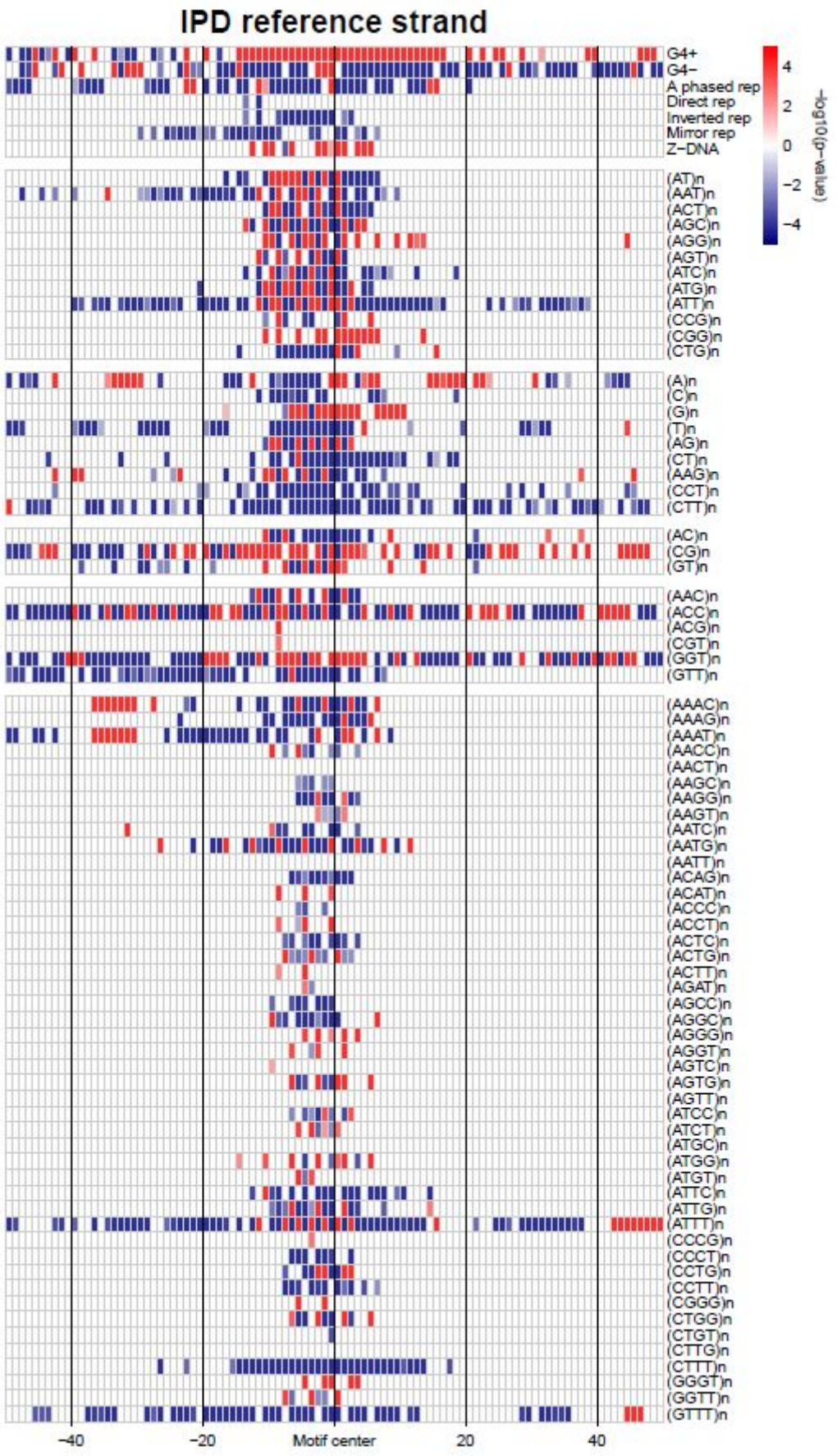
A


IPD reference strand
G4+
G4−
A phased rep
Direct rep
Inverted rep
Mirror rep
Z−DNA
(AT)n
(AAT)n
(ACT)n
(AGC)n
(AGG)n
(AGT)n
(ATC)n
(ATG)n
(ATT)n
(CCG)n
(CGG)n
(CTG)n
(A)n
(C)n
(G)n
(T)n
(AG)n
(CT)n
(AAG)n
(CCT)n
(CTT)n
(AC)n
(CG)n
(GT)n
(AAC)n
(ACC)n
(ACG)n
(CGT)n
(GGT)n
(GTT)n
(AAAC)n
(AAAG)n
(AAAT)n
(AACC)n
(AACT)n
(AAGC)n
(AAGG)n
(AAGT)n
(AATC)n
(AATG)n
(AATT)n
(ACAG)n
(ACAT)n
(ACCC)n
(ACCT)n
(ACTC)n
(ACTG)n
(ACTT)n
(AGAT)n
(AGCC)n
(AGGC)n
(AGGG)n
(AGGT)n
(AGTC)n
(AGTG)n
(AGTT)n
(ATCC)n
(ATCT)n
(ATGC)n
(ATGG)n
(ATGT)n
(ATTC)n
(ATTG)n
(ATTT)n
(CCCG)n
(CCCT)n
(CCTG)n
(CCTT)n
(CGGG)n
(CTGG)n
(CTGT)n
(CTTG)n
(CTTT)n
(GGGT)n
(GGTT)n
(GTTT)n
4
2
0
−2
−4
−log10(p−value)
−40
−20
Motif center
20
40

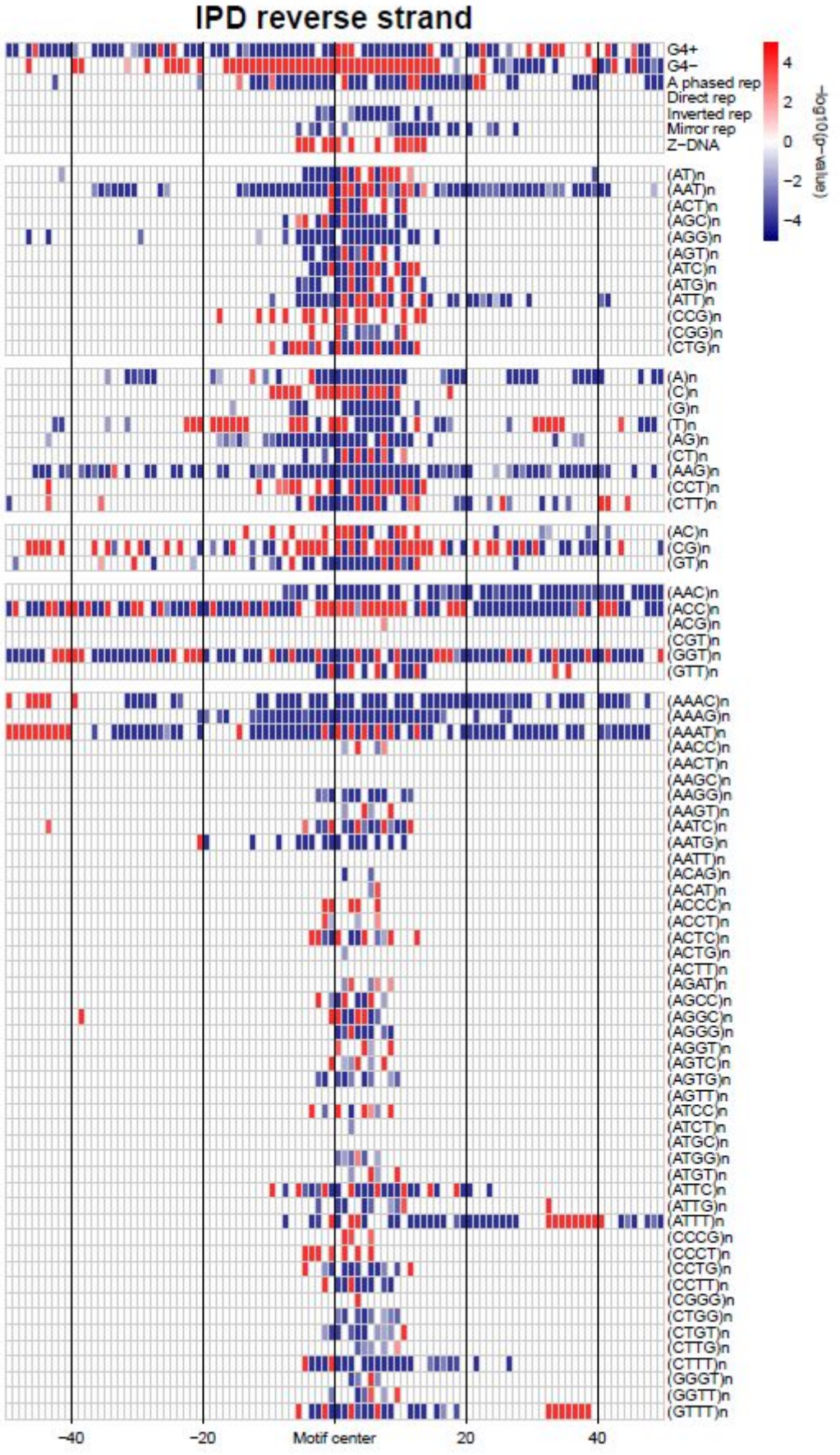
B


IPD reverse strand
G4+
G4−
A phased rep
Direct rep
Inverted rep
Mirror rep
Z−DNA
(AT)n
(AAT)n
(ACT)n
(AGC)n
(AGG)n
(AGT)n
(ATC)n
(ATG)n
(ATT)n
(CCG)n
(CGG)n
(CTG)n
(A)n
(C)n
(G)n
(T)n
(AG)n
(CT)n
(AAG)n
(CCT)n
(CTT)n
(AC)n
(CG)n
(GT)n
(AAC)n
(ACC)n
(ACG)n
(CGT)n
(GGT)n
(GTT)n
(AAAC)n
(AAAG)n
(AAAT)n
(AACC)n
(AACT)n
(AAGC)n
(AAGG)n
(AAGT)n
(AATC)n
(AATG)n
(AATT)n
(ACAG)n
(ACAT)n
(ACCC)n
(ACCT)n
(ACTC)n
(ACTG)n
(ACTT)n
(AGAT)n
(AGCC)n
(AGGC)n
(AGGG)n
(AGGT)n
(AGTC)n
(AGTG)n
(AGTT)n
(ATCC)n
(ATCT)n
(ATGC)n
(ATGG)n
(ATGT)n
(ATTC)n
(ATTG)n
(ATTT)n
(CCCG)n
(CCCT)n
(CCTG)n
(CCTT)n
(CGGG)n
(CTGG)n
(CTGT)n
(CTTG)n
(CTTT)n
(GGGT)n
(GGTT)n
(GTTT)n
4
2
0
−2
−4
−log10(p−value)
−40
−20
Motif center
20
40

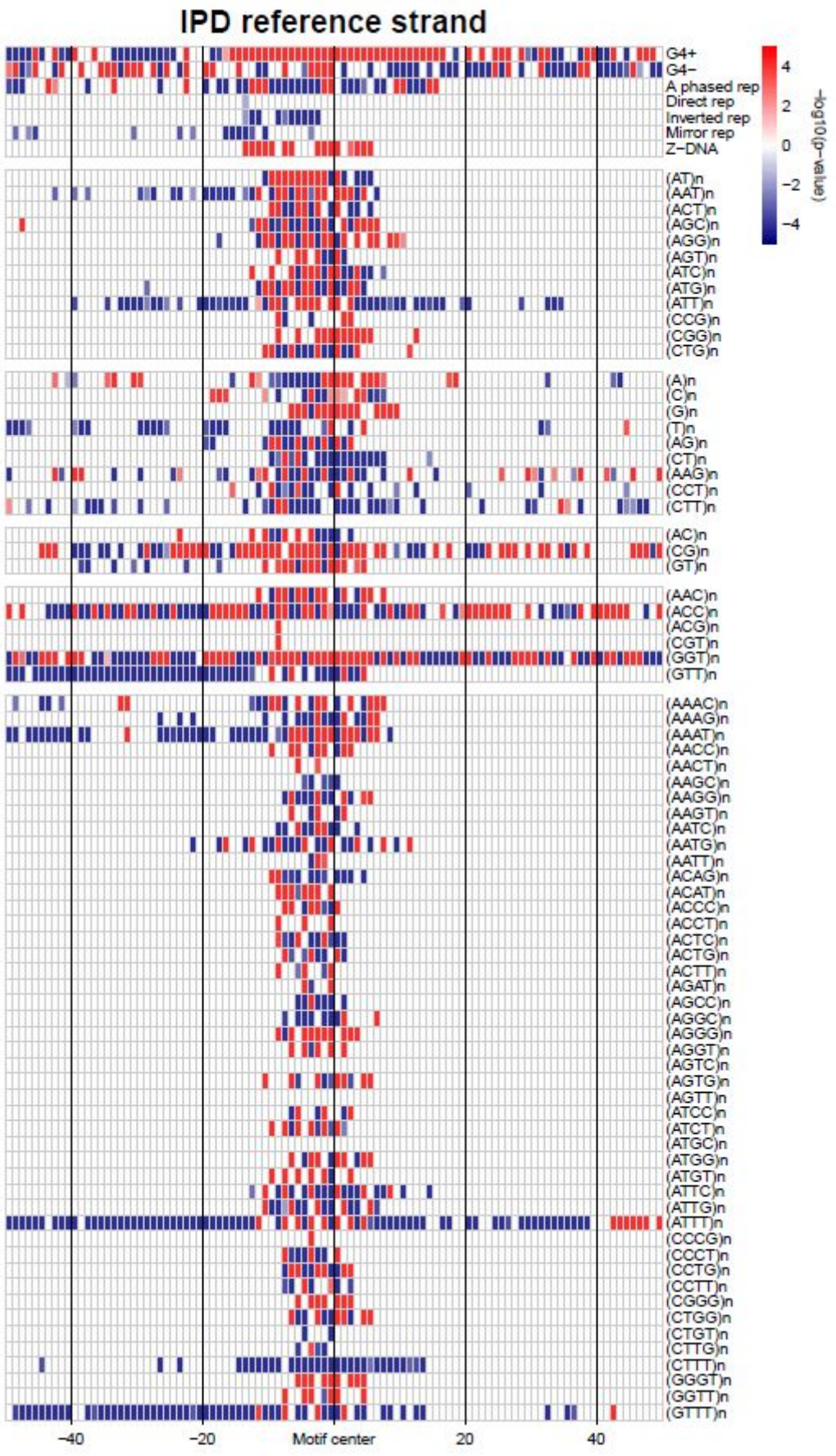
C


IPD reference strand
G4+
G4−
A phased rep
Direct rep
Inverted rep
Mirror rep
Z−DNA
(AT)n
(AAT)n
(ACT)n
(AGC)n
(AGG)n
(AGT)n
(ATC)n
(ATG)n
(ATT)n
(CCG)n
(CGG)n
(CTG)n
(A)n
(C)n
(G)n
(T)n
(AG)n
(CT)n
(AAG)n
(CCT)n
(CTT)n
(AC)n
(CG)n
(GT)n
(AAC)n
(ACC)n
(ACG)n
(CGT)n
(GGT)n
(GTT)n
(AAAC)n
(AAAG)n
(AAAT)n
(AACC)n
(AACT)n
(AAGC)n
(AAGG)n
(AAGT)n
(AATC)n
(AATG)n
(AATT)n
(ACAG)n
(ACAT)n
(ACCC)n
(ACCT)n
(ACTC)n
(ACTG)n
(ACTT)n
(AGAT)n
(AGCC)n
(AGGC)n
(AGGG)n
(AGGT)n
(AGTC)n
(AGTG)n
(AGTT)n
(ATCC)n
(ATCT)n
(ATGC)n
(ATGG)n
(ATGT)n
(ATTC)n
(ATTG)n
(ATTT)n
(CCCG)n
(CCCT)n
(CCTG)n
(CCTT)n
(CGGG)n
(CTGG)n
(CTGT)n
(CTTG)n
(CTTT)n
(GGGT)n
(GGTT)n
(GTTT)n
4
2
0
−2
−4
−log10(p−value)
−40
−20
Motif center
20
40

D


IPD reverse strand
G4+
G4−
A phased rep
Direct rep
Inverted rep
Mirror rep
Z−DNA
(AT)n
(AAT)n
(ACT)n
(AGC)n
(AGG)n
(AGT)n
(ATC)n
(ATG)n
(ATT)n
(CCG)n
(CGG)n
(CTG)n
(A)n
(C)n
(G)n
(T)n
(AG)n
(CT)n
(AAG)n
(CCT)n
(CTT)n
(AC)n
(CG)n
(GT)n
(AAC)n
(ACC)n
(ACG)n
(CGT)n
(GGT)n
(GTT)n
(AAAC)n
(AAAG)n
(AAAT)n
(AACC)n
(AACT)n
(AAGC)n
(AAGG)n
(AAGT)n
(AATC)n
(AATG)n
(AATT)n
(ACAG)n
(ACAT)n
(ACCC)n
(ACCT)n
(ACTC)n
(ACTG)n
(ACTT)n
(AGAT)n
(AGCC)n
(AGGC)n
(AGGG)n
(AGGT)n
(AGTC)n
(AGTG)n
(AGTT)n
(ATCC)n
(ATCT)n
(ATGC)n
(ATGG)n
(ATGT)n
(ATTC)n
(ATTG)n
(ATTT)n
(CCCG)n
(CCCT)n
(CCTG)n
(CCTT)n
(CGGG)n
(CTGG)n
(CTGT)n
(CTTG)n
(CTTT)n
(GGGT)n
(GGTT)n
(GTTT)n
4
2
0
−2
−4
−log10(p−value)
−40
−20
Motif center
20
40

E


IPD reference strand
G4+
G4−
A phased rep
Direct rep
Inverted rep
Mirror rep
Z−DNA
(AT)n
(AAT)n
(ACT)n
(AGC)n
(AGG)n
(AGT)n
(ATC)n
(ATG)n
(ATT)n
(CCG)n
(CGG)n
(CTG)n
(A)n
(C)n
(G)n
(T)n
(AG)n
(CT)n
(AAG)n
(CCT)n
(CTT)n
(AC)n
(CG)n
(GT)n
(AAC)n
(ACC)n
(ACG)n
(CGT)n
(GGT)n
(GTT)n
(AAAC)n
(AAAG)n
(AAAT)n
(AACC)n
(AACT)n
(AAGC)n
(AAGG)n
(AAGT)n
(AATC)n
(AATG)n
(AATT)n
(ACAG)n
(ACAT)n
(ACCC)n
(ACCT)n
(ACTC)n
(ACTG)n
(ACTT)n
(AGAT)n
(AGCC)n
(AGGC)n
(AGGG)n
(AGGT)n
(AGTC)n
(AGTG)n
(AGTT)n
(ATCC)n
(ATCT)n
(ATGC)n
(ATGG)n
(ATGT)n
(ATTC)n
(ATTG)n
(ATTT)n
(CCCG)n
(CCCT)n
(CCTG)n
(CCTT)n
(CGGG)n
(CTGG)n
(CTGT)n
(CTTG)n
(CTTT)n
(GGGT)n
(GGTT)n
(GTTT)n
4
2
0
−2
−4
−log10(p−value)
−40
−20
Motif center
20
40

F

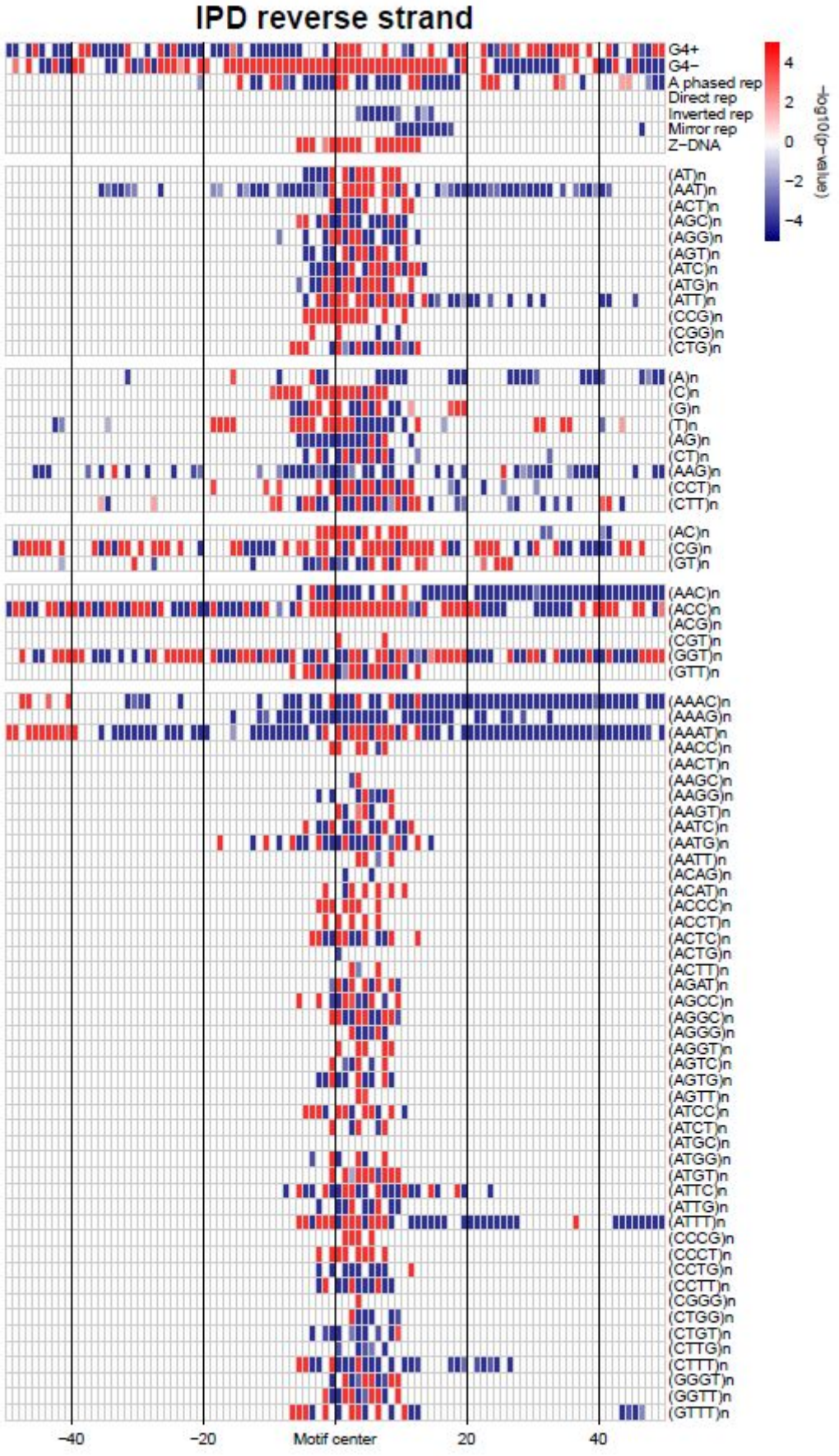
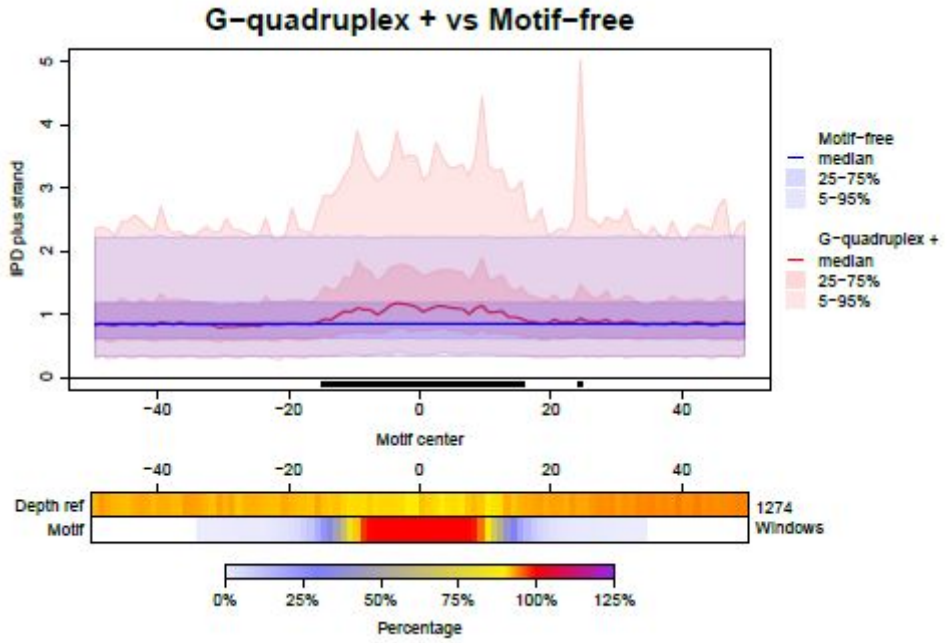
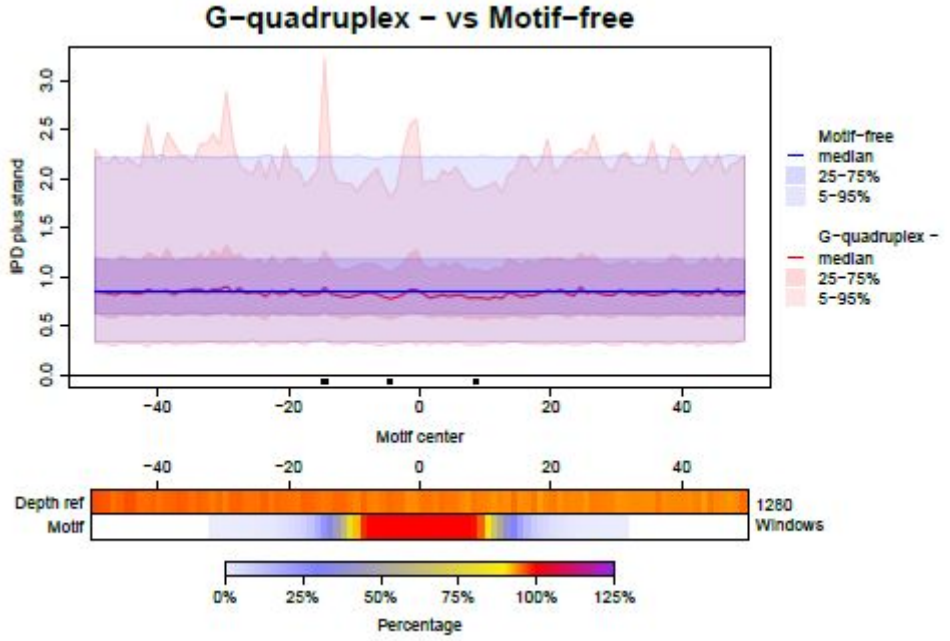

IPD reverse strand
G4+
G4−
A phased rep
Direct rep
Inverted rep
Mirror rep
Z−DNA
(AT)n
(AAT)n
(ACT)n
(AGC)n
(AGG)n
(AGT)n
(ATC)n
(ATG)n
(ATT)n
(CCG)n
(CGG)n
(CTG)n
(A)n
(C)n
(G)n
(T)n
(AG)n
(CT)n
(AAG)n
(CCT)n
(CTT)n
(AC)n
(CG)n
(GT)n
(AAC)n
(ACC)n
(ACG)n
(CGT)n
(GGT)n
(GTT)n
(AAAC)n
(AAAG)n
(AAAT)n
(AACC)n
(AACT)n
(AAGC)n
(AAGG)n
(AAGT)n
(AATC)n
(AATG)n
(AATT)n
(ACAG)n
(ACAT)n
(ACCC)n
(ACCT)n
(ACTC)n
(ACTG)n
(ACTT)n
(AGAT)n
(AGCC)n
(AGGC)n
(AGGG)n
(AGGT)n
(AGTC)n
(AGTG)n
(AGTT)n
(ATCC)n
(ATCT)n
(ATGC)n
(ATGG)n
(ATGT)n
(ATTC)n
(ATTG)n
(ATTT)n
(CCCG)n
(CCCT)n
(CCTG)n
(CCTT)n
(CGGG)n
(CTGG)n
(CTGT)n
(CTTG)n
(CTTT)n
(GGGT)n
(GGTT)n
(GTTT)n
4
2
0
−2
−4
−log10(p−value)
−40
−20
Motif center
20
40

Figure S12. Variation in IPD remains in PCR-amplified sequences.
The chromosome 21 from Sumatran orangutan was flow-sorted from a cell line using a previously described protocol(*116*). Subsequently, the flow-sorted material was used as a template for WGA performed with the REPLI-g Single Cell Kit (Qiagen). After de-branching(*117*), the whole-genome amplified material was sequenced on 4 SMRT cells of the RSII instrument. Non-B DNA annotations of orangutan were obtained from the non-B DB (*110*). **A** G+ motifs. **B** G- motifs. **C** A-phased repeats. **D** Direct repeats. **E** Inverted repeats. **F** Mirror repeats. **G** Z-DNA motifs. See the legend of Fig. 2A for details.

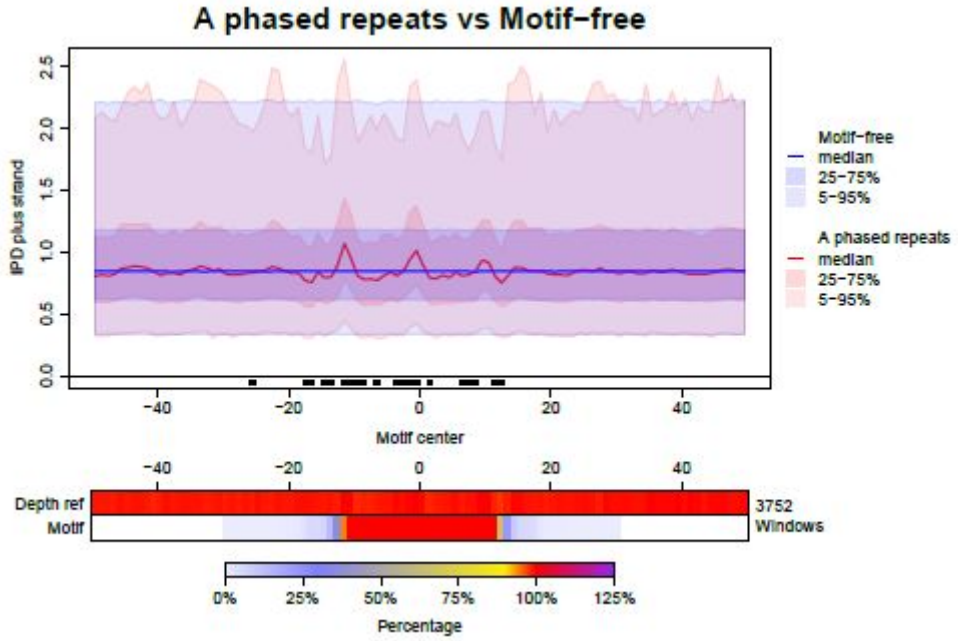
A



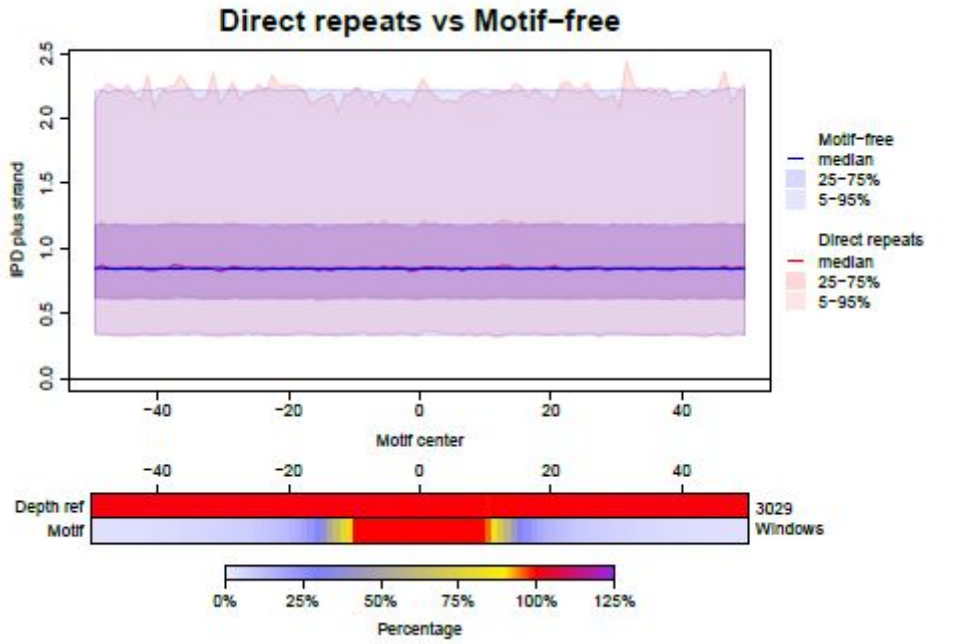
B



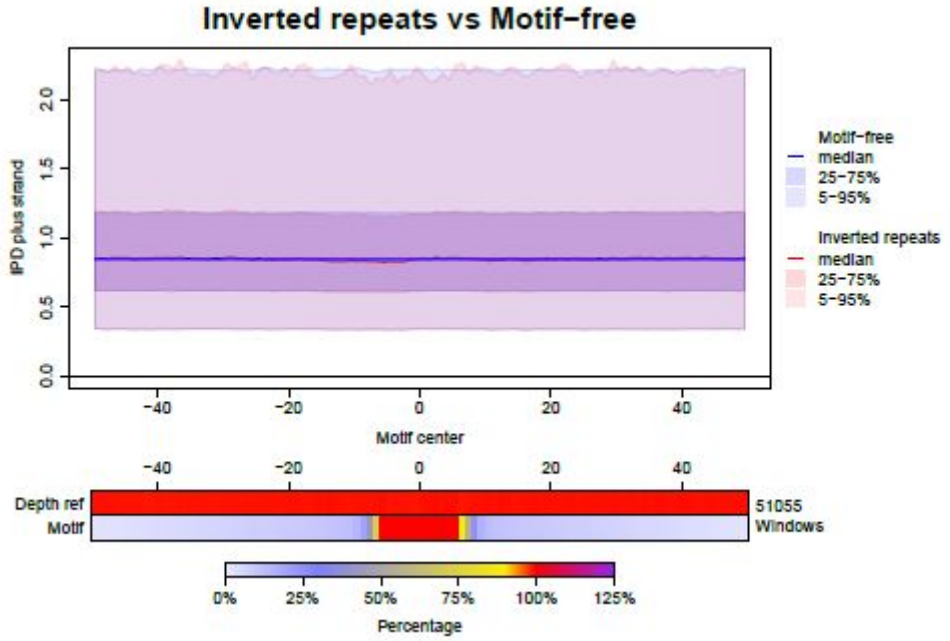
C


A phased repeats vs Motif−free
IPD plus strand
0.0
0.5
1.0
1.5
2.0
2.5
−40
−20
0
20
40
Motif center
Motif−free
median
25−75%
5−95%
A phased repeats
median
25−75%
5−95%
Depth ref
Motif
3752
Windows
0%
25%
50%
75%
100%
125%
Percentage

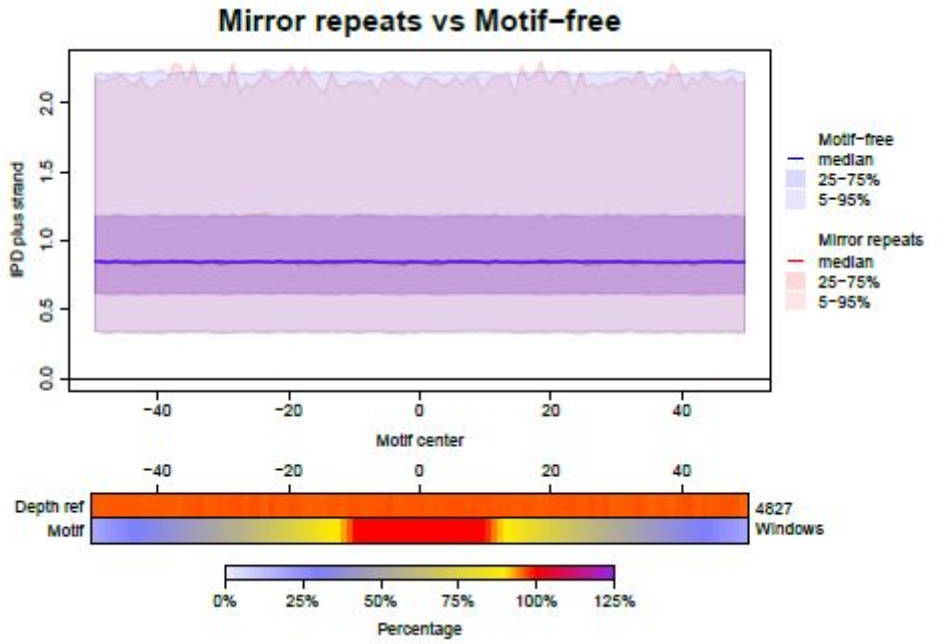
D


Direct repeats vs Motif−free
IPD plus strand
0.0
0.5
1.0
1.5
2.0
2.5
−40
−20
0
20
40
Motif center
Motif−free
median
25−75%
5−95%
Direct repeats
median
25−75%
5−95%
Depth ref
Motif
3029
Windows
0%
25%
50%
75%
100%
125%
Percentage

E


Inverted repeats vs Motif−free
IPD plus strand
0.0
0.5
1.0
1.5
2.0
−40
−20
0
20
40
Motif center
Motif−free
median
25−75%
5−95%
Inverted repeats
median
25−75%
5−95%
Depth ref
Motif
51055
Windows
0%
25%
50%
75%
100%
125%
Percentage

F


Mirror repeats vs Motif−free
IPD plus strand
0.0
0.5
1.0
1.5
2.0
−40
−20
0
20
40
Motif center
Motif−free
median
25−75%
5−95%
Mirror repeats
median
25−75%
5−95%
Depth ref
Motif
4827
Windows
0%
25%
50%
75%
100%
125%
Percentage

G

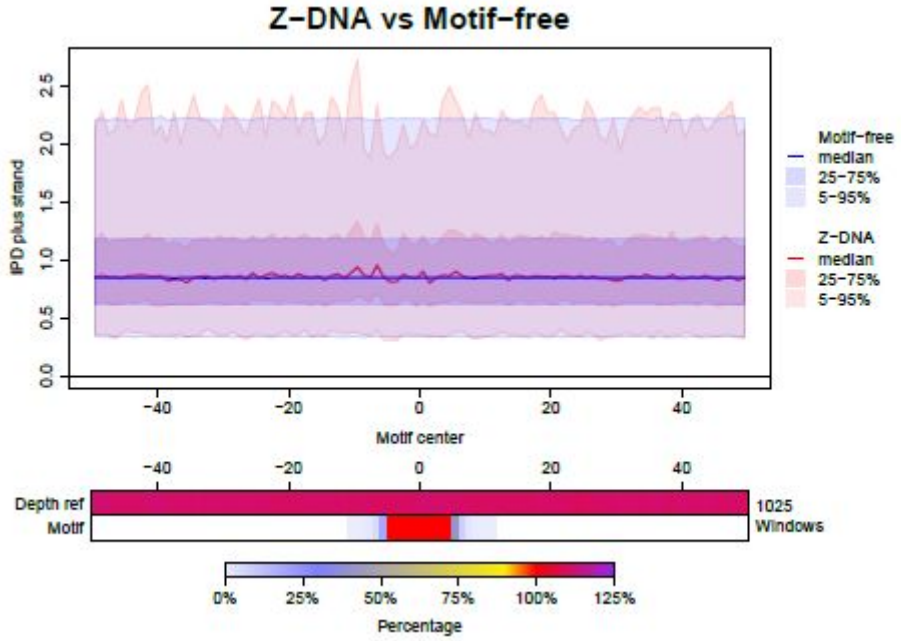

Z−DNA vs Motif−free
IPD plus strand
0.0
0.5
1.0
1.5
2.0
2.5
−40
−20
0
20
40
Motif center
Motif−free
median
25−75%
5−95%
Z−DNA
median
25−75%
5−95%
Depth ref
Motif
1025
Windows
0%
25%
50%
75%
100%
125%
Percentage

Figure S13. The relationship between IPD and sequence composition.
Plot of the mean IPD in each motif-free window in relation to sequence composition (percentage of A, T, G and C in the window). The red clouds indicate observed IPDs, while the blue clouds correspond to the compositional regression model with the mean IPD as response and the single nucleotide sequence composition as the predictor. The top right of each panel reports the correlation between the percentage of each nucleotide and the mean IPD in motif-free windows.

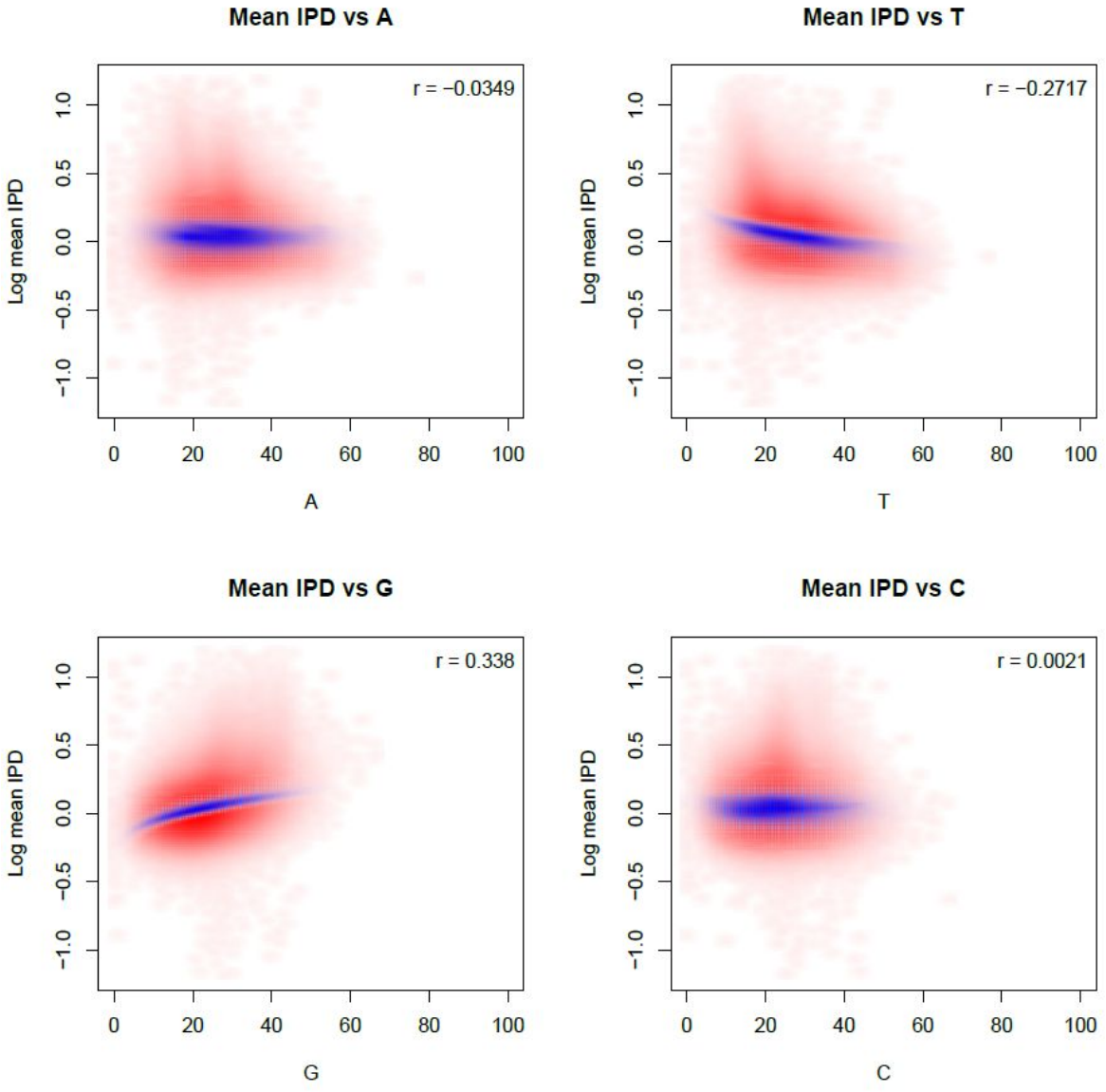
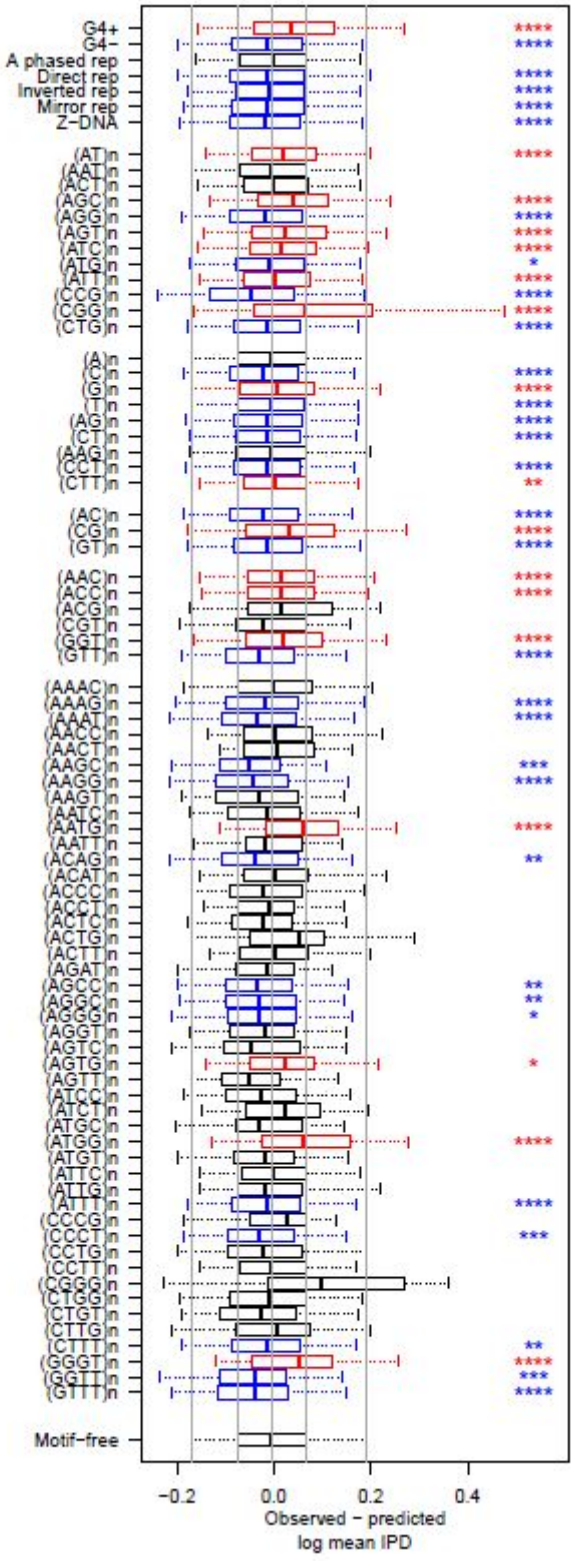


Figure S14. A comparison between observed and predicted mean IPD.
Predictions of mean IPD values in motif-containing windows are obtained from a compositional regression model fitted considering dinucleotide sequence composition on motif-free windows. **A** Reference strand. **B** Reverse complement strand. See the legend of Fig. 2F for details.

A



B

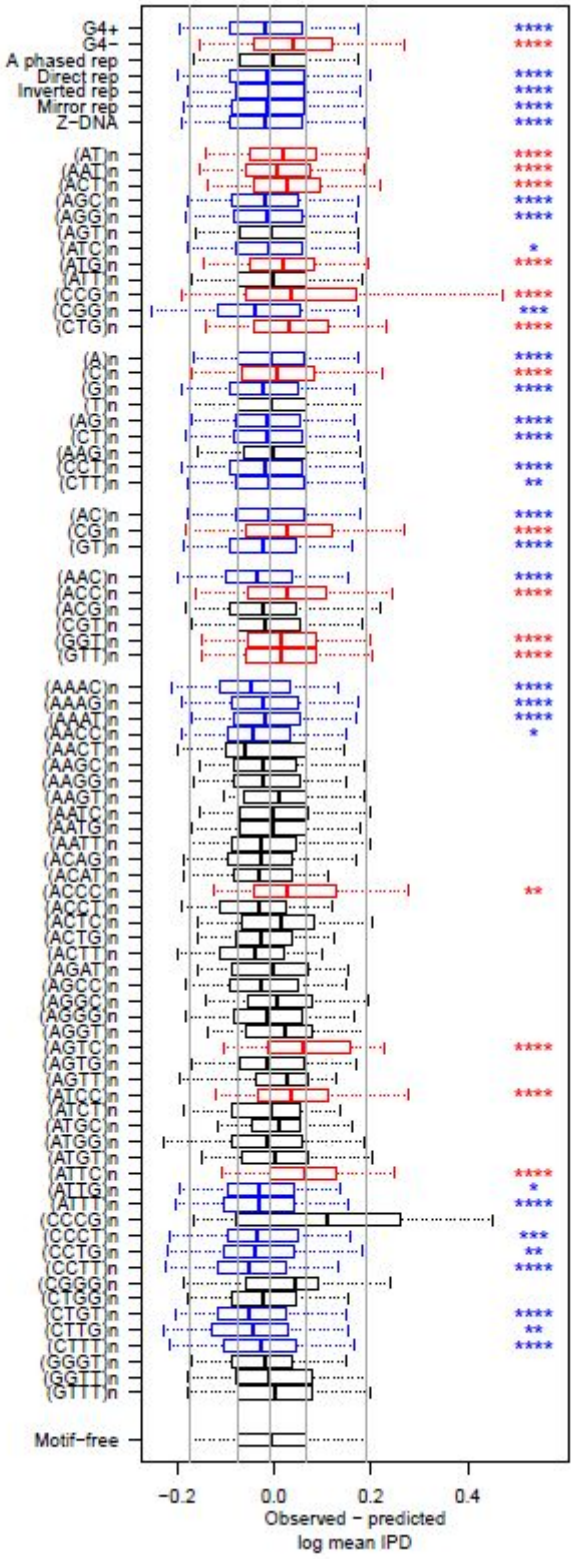
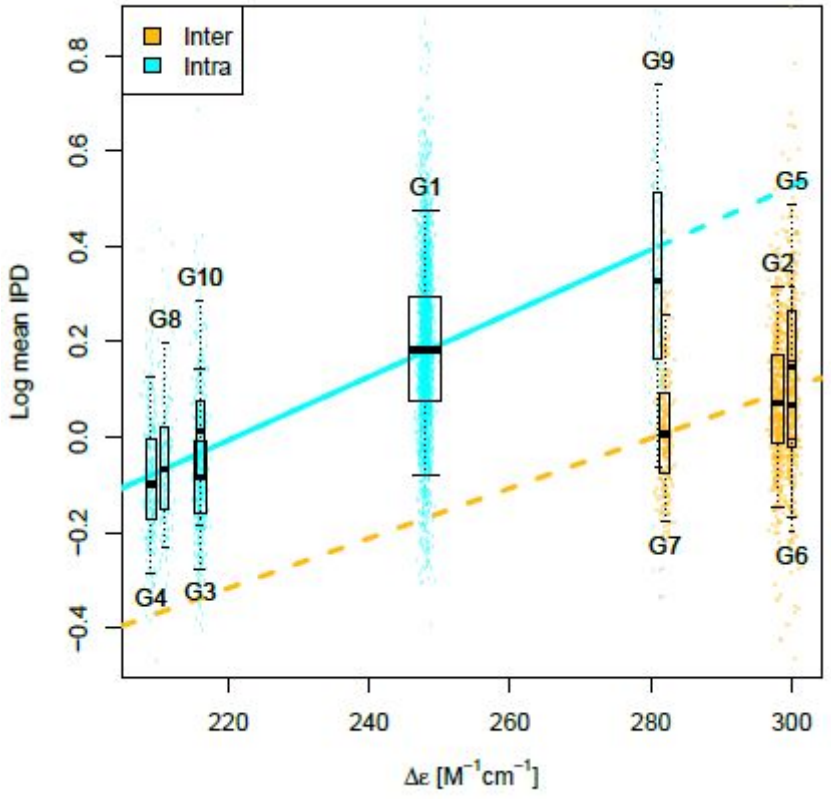

G4+
G4−
A phased rep
Direct rep
Inverted rep
Mirror rep
Z−DNA
(AT)n
(AAT)n
(ACT)n
(AGC)n
(AGG)n
(AGT)n
(ATC)n
(ATG)n
(ATT)n
(CCG)n
(CGG)n
(CTG)n
(A)n
(C)n
(G)n
(T)n
(AG)n
(CT)n
(AAG)n
(CCT)n
(CTT)n
(AC)n
(CG)n
(GT)n
(AAC)n
(ACC)n
(ACG)n
(CGT)n
(GGT)n
(GTT)n
(AAAC)n
(AAAG)n
(AAAT)n
(AACC)n
(AACT)n
(AAGC)n
(AAGG)n
(AAGT)n
(AATC)n
(AATG)n
(AATT)n
(ACAG)n
(ACAT)n
(ACCC)n
(ACCT)n
(ACTC)n
(ACTG)n
(ACTT)n
(AGAT)n
(AGCC)n
(AGGC)n
(AGGG)n
(AGGT)n
(AGTC)n
(AGTG)n
(AGTT)n
(ATCC)n
(ATCT)n
(ATGC)n
(ATGG)n
(ATGT)n
(ATTC)n
(ATTG)n
(ATTT)n
(CCCG)n
(CCCT)n
(CCTG)n
(CCTT)n
(CGGG)n
(CTGG)n
(CTGT)n
(CTTG)n
(CTTT)n
(GGGT)n
(GGTT)n
(GTTT)n
Motif−free
−0.2
0.0
0.2
0.4
Observed − predicted
log mean IPD

Figure S15. G-quadruplex thermostability and molecularity as predictors of polymerization kinetics.

G1 through G10 indicate, in order, the ten most common G-quadruplex motif types in our annotations (G1 the most common, G2 the next most common, etc.; Table S5). For each motif type we measured delta epsilon and T_m once, while we computed an average IPD for each occurrence of the motif in the genome, thus thousands of motifs were analyzed (Table S5). The average IPD value was then regressed against **A** circular dichroism (delta epsilon), or **B** light absorption (melting temperature, T_m), considering intra- and intermolecular G4s together and using molecularity (intra/inter-strandedness) as a binary predictor (dashed lines; solid lines represent the model in Fig. 3 obtained using only intramolecular G4s). R-squared 28.4% for delta epsilon (molecularity significantly changes the slope, but not the intercept, of the line), 6.7% for T_m (molecularity significantly changes both the slope and the intercept of the line). Yellow: intermolecular G-quadruplexes. Cyan: intramolecular G-quadruplexes. Boxplot whiskers mark the 5th and 95th quantiles. R-squared 28.4% for delta epsilon (molecularity significantly changes the slope, but not the intercept, of the line), 6.7% for T_m (molecularity significantly changes both the slope and the intercept of the line). Yellow: intermolecular G-quadruplexes. Cyan: intramolecular G-quadruplexes. Boxplot whiskers mark the 5th and 95th quantiles.

A



B

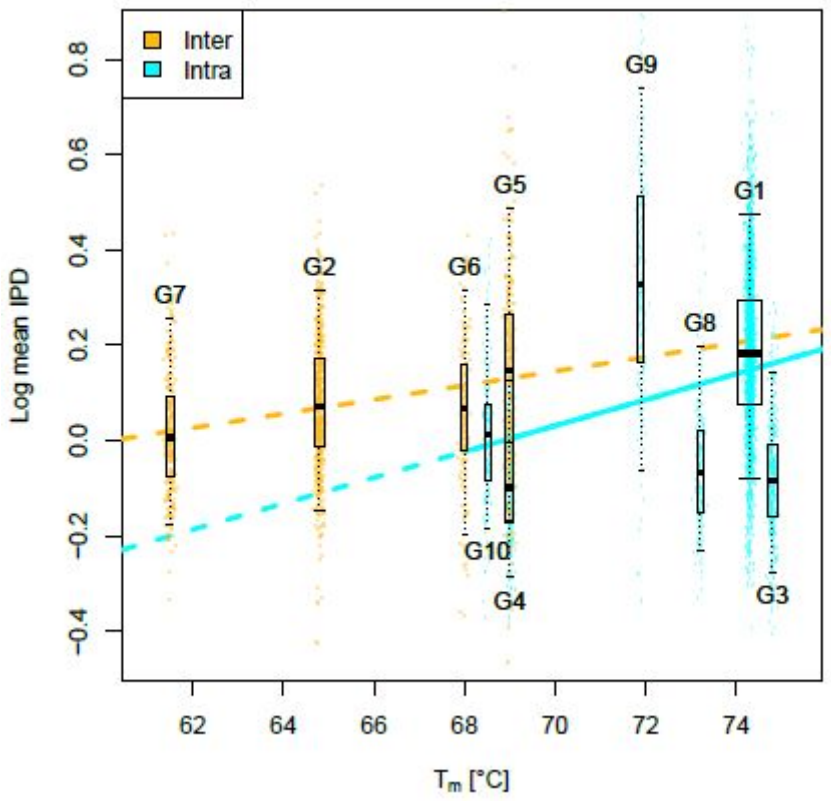

Inter
Intra
Log mean IPD
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
G7
G2
G6
G5
G9
G1
G8
G10
G4
G3
62
64
66
68
70
72
74
T_m [°C]

Figure S16. CD spectra, thermal denaturation and PAGE.

A $(GGT)_4$. **B** $(GGT)_5$. **C** $(GGT)_6$. CD spectra of all three oligonucleotides were measured at various potassium concentrations and kinetics (after 30 minutes period after K+ addition or after slow annealing). Insert figures show thermal denaturation curves and T_m. (D) Native 16% PAGE (10mM K-phosphate+35mM KCl, pH 7.0, stained by Stains All) shows tetramolecular quadruplex in $(GGT)_4$, bimolecular quadruplex in $(GGT)_5$ and bi- and monomolecular quadruplex in $(GGT)_6$. Samples in the PAGE were slowly annealed for 2 hours before loading onto the gel.

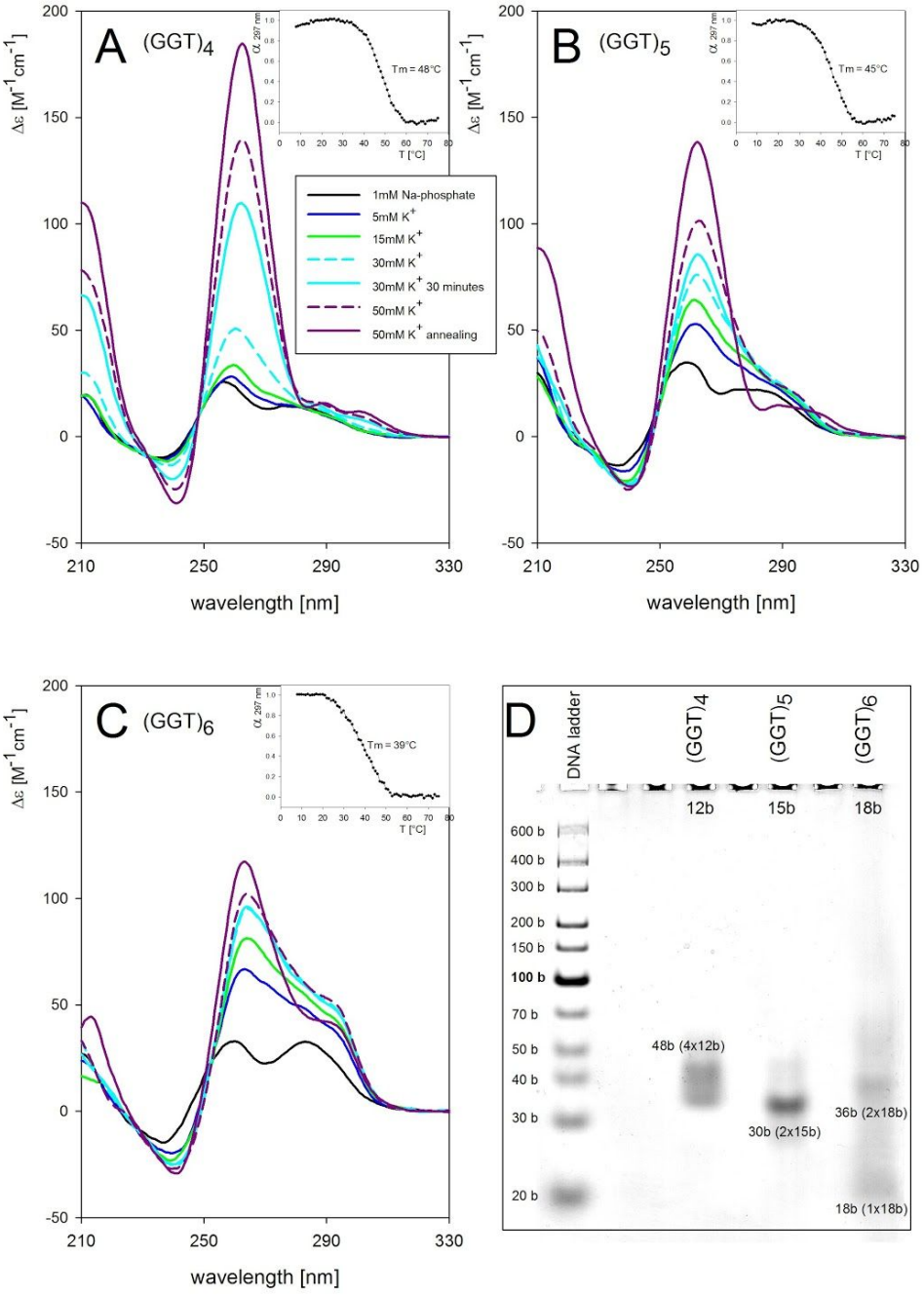


Figure S17. Effects of non-B DNA motifs on insertions as sequencing errors or mutations. See legend from Fig. 4.

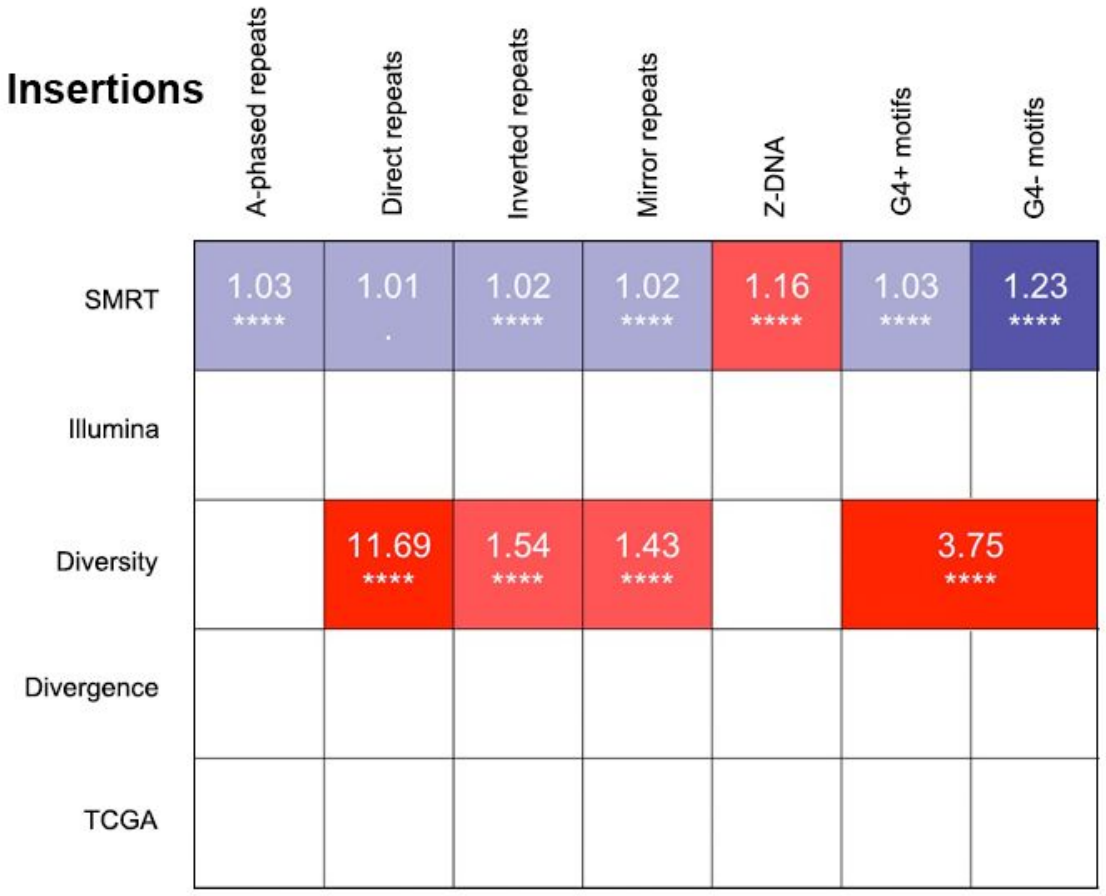
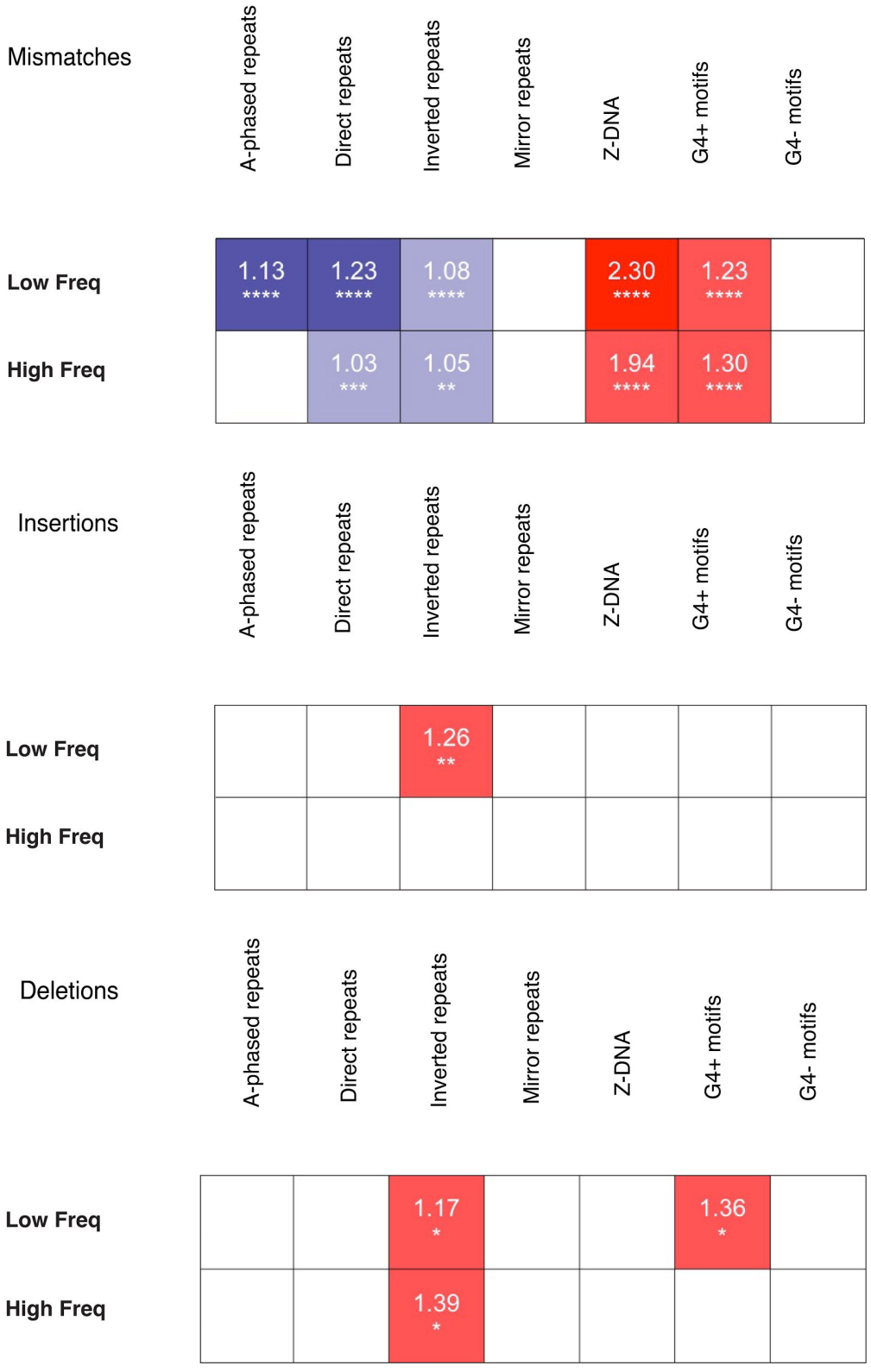


Figure S18. Effects of non-B DNA motifs on low (minor allele frequency between 1% and 5%) and high (minor allele frequency above 5%) frequency variants in the 1,000 Genomes Project. See Figure 4 legend for details.



REFERENCES

83. G. Ananda *et al.*, Distinct mutational behaviors differentiate short tandem repeats from microsatellites in the human genome. *Genome Biol. Evol.* 5, 606–620 (2013).

84. A. Rhoads, K. F. Au, PacBio Sequencing and Its Applications. *Genomics Proteomics Bioinformatics*. 13, 278–289 (2015).

85. R. Campos-Sánchez, M. A. Cremona, A. Pini, F. Chiaromonte, K. D. Makova, Integration and Fixation Preferences of Human and Mouse Endogenous Retroviruses Uncovered with Functional Data Analysis. *PLoS Comput. Biol.* 12, e1004956 (2016).

86. V. Pawlowsky-Glahn, J. J. Egozcue, R. Tolosana-Delgado, *Modeling and Analysis of Compositional Data* (John Wiley & Sons, 2015).

87. I. Kejnovská *et al.*, Clustered abasic lesions profoundly change the structure and stability of human telomeric G-quadruplexes. *Nucleic Acids Res.* 45, 4294–4305 (2017).

88. D. Blankenberg *et al.*, Dissemination of scientific software with Galaxy ToolShed. *Genome Biol.* 15, 403 (2014).

89. M. Blanchette *et al.*, Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.* 14, 708–715 (2004).

90. R. S. Harris, thesis, Pennsylvania State University (2007).

91. S. B. Montgomery *et al.*, The origin, evolution, and functional impact of short insertion–deletion variants identified in 179 human genomes. *Genome Res.* 23, 749–761 (2013).

92. K. Cibulskis *et al.*, Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* 31, 213–219 (2013).

93. S. Taylor, K. Pollard, *Stat. Appl. Genet. Mol. Biol.*, in press.

94. P. A. Lachenbruch, Analysis of data with clumping at zero. *Biom. Z.* 18, 351–356 (1976).

95. M. Schirmer, R. D'Amore, U. Z. Ijaz, N. Hall, C. Quince, Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. *BMC Bioinformatics*. 17, 125 (2016).

96. H. Li, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv [q-bio.GN]* (2013), (available at http://arxiv.org/abs/1303.3997).

97. M. J. Chaisson, G. Tesler, Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics*. 13, 238 (2012).

98. C. Hercus, Novoalign. *Selangor: Novocraft Technologies* (2012).

99. R. S. Harris, *Improved pairwise alignment of genomic DNA* (The Pennsylvania State University, 2007).

100. W. Miller *et al.*, 28-way vertebrate alignment and conservation track in the UCSC Genome Browser. *Genome Res.* 17, 1797–1808 (2007).

101. J. M. Zook *et al.*, Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci Data*. 3, 160025 (2016).

102. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat. Methods*. 9, 357–359 (2012).

103. S. M. Kiełbasa, R. Wan, K. Sato, P. Horton, M. C. Frith, Adaptive seeds tame genomic sequence comparison. *Genome Res.* 21, 487–493 (2011).

104. G. Lunter, M. Goodson, Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res.* 21, 936–939 (2011).

105. A. Pini, S. Vantini, The interval testing procedure: A general framework for inference in functional data analysis. *Biometrics*. 72, 835–845 (2016).

106. O. Vsevolozhskaya, M. Greenwood, D. Holodov, Pairwise comparison of treatment levels in functional analysis of variance with application to erythrocyte hemolysis. *Ann. Appl. Stat.* 8, 905–925 (2014).

107. R. Campos-Sánchez, M. A. Cremona, A. Pini, F. Chiaromonte, K. D. Makova, Integration and Fixation Preferences of Human and Mouse Endogenous Retroviruses Uncovered with Functional Data Analysis. *PLoS Comput. Biol.* 12, e1004956 (2016).

108. Cremona MA, Pini A, Cumbo F, Makova KD, Chiaromonte F and Vantini S, IWTomics: testing high resolution “Omics” data at multiple locations and scales. *Submitted.* (2017).

109. Pini A, Vantini S, Interval-wise testing for functional data. *J. Nonparametr. Stat.* 29, 407–424 (2017).

110. R. Z. Cer *et al.*, Non-B DB v2.0: a database of predicted non-B DNA-forming motifs and its associated tools. *Nucleic Acids Res.* 41, D94–D100 (2013).

111. A. Fungtammasan *et al.*, Accurate typing of short tandem repeats from genome-wide

sequencing data and its applications. *Genome Res.* 25, 736–749 (2015).

112. R. R. Sinden, *DNA Structure and Function* (Elsevier, 2012).

113. E. V. Mirkin, S. M. Mirkin, Replication fork stalling at natural impediments. *Microbiol. Mol. Biol. Rev.* 71, 13–35 (2007).

114. D. Huertas, F. Azorín, Structural polymorphism of homopurine DNA sequences. d(GGA)n and d(GGGA)n repeats form intramolecular hairpins stabilized by different base-pairing interactions. *Biochemistry*. 35, 13125–13135 (1996).

115. E. Trotta, N. Del Grosso, M. Erba, M. Paci, The ATT Strand of AAT⊙ ATT Trinucleotide Repeats Adopts Stable Hairpin Structures Induced by Minor Groove Binding Ligands. *Biochemistry*. 39, 6799–6808 (2000).

116. F. Yang, N. P. Carter, L. Shi, M. A. Ferguson-Smith, A comparative study of karyotypes of muntjacs by chromosome painting. *Chromosoma*. 103, 642–652 (1995).

117. K. Zhang *et al.*, Sequencing genomes from single cells by polymerase cloning. *Nat. Biotechnol.* 24, 680–686 (2006).