SUPPLEMENTARY INFORMATION

SUPPLEMENTARY TEXT

Interval-Wise Testing: statistical details.

The IWT is a novel inferential procedure for functional data (Pini and Vantini 2017) 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 (Pini and Vantini 2016; Vsevolozhskaya et al. 2014) 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) (Pini and Vantini 2016) was based on a basis expansion of the curves, while the procedure developed in (Vsevolozhskaya et al. 2014) 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 (Campos-Sánchez et al. 2016) 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 (Cremona et al. 2018). 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 $IPD_{f,i}(t)\ i = 1,\dots,n_f$ be the IPD curves in the n_f motif-containing windows (features), and $IPD_{c,i}(t)\ i = 1,\dots,n_c$ 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 $IPD_{f,i}(t) i = 1,\dots,n_f$ and $IPD_{c,i}(t)\ i = 1,\dots,n_c$ 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 $\overline{IPD_c}(t) = \frac{1}{n_c}\sum_{i=1}^{n_c} IPD_{ci}(t)$ 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

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

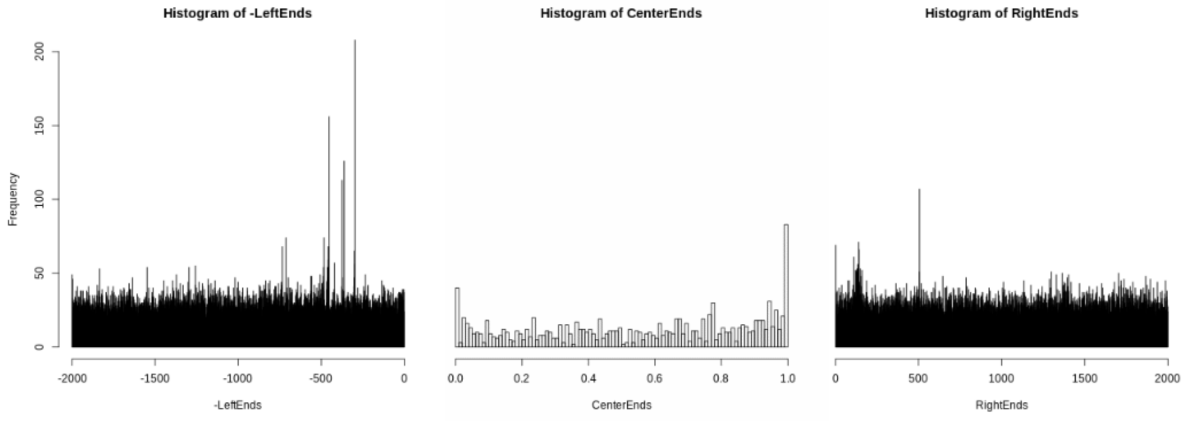
This multiple testing correction controls the interval-wise error rate; that is, $\tilde{p}(t)$ controls the probability of rejecting the null hypothesis H_0^S on every interval $S \subseteq I$where it is true (see details in(Pini A, Vantini S 2017)). Finally, we identify locations with a significant 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 $\tilde{p}_s(t) \leq \alpha$).

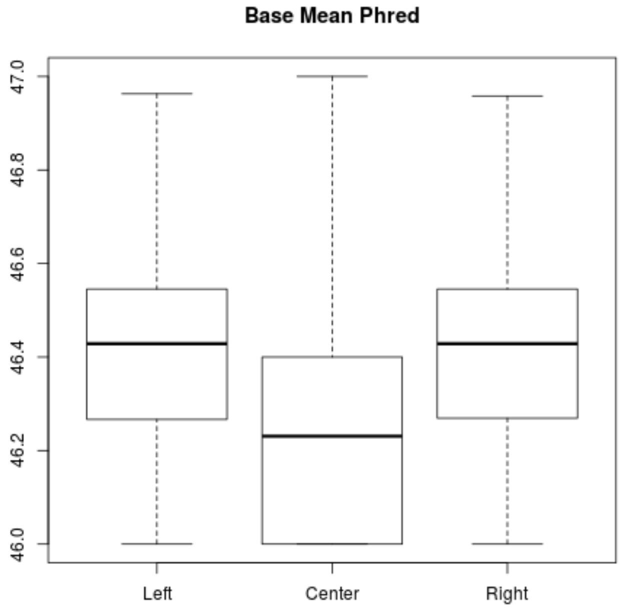
SUPPLEMENTARY NOTE 1

Analysis of sequencing depth decrease at G4 motifs

To explore the causes behind lower sequencing depth at G4 motifs, we performed two analyses. First, we investigated whether read terminations were more common in G4 motifs than in their flanks. We recovered the ending positions of reads aligned to chromosome 21 using the pysam function reference_end. Coordinates were intersected with annotated G4+ (CenterEnds) and 2-kb windows upstream (LeftEnds) and downstream (RightEnds). Because of the variable size of G4+ motifs, the coordinates of read ends intersecting with a motif were scaled to map in a [0,1] interval. We observed a decreased amount of read ends in G4 motifs compared to their flanks. Thus, reads do not appear to preferentially end within G4 motifs.



Second, we investigated whether lower depth might be due to lower sequencing quality (PHRED scores). Mean sequencing qualities (per base) were retrieved from all reads mapping to coordinates intersecting G4+ motifs and their 2kb flanks. Sequencing qualities in motifs were lower than in their flanks (*t*-test p-value < 2.2e-16).

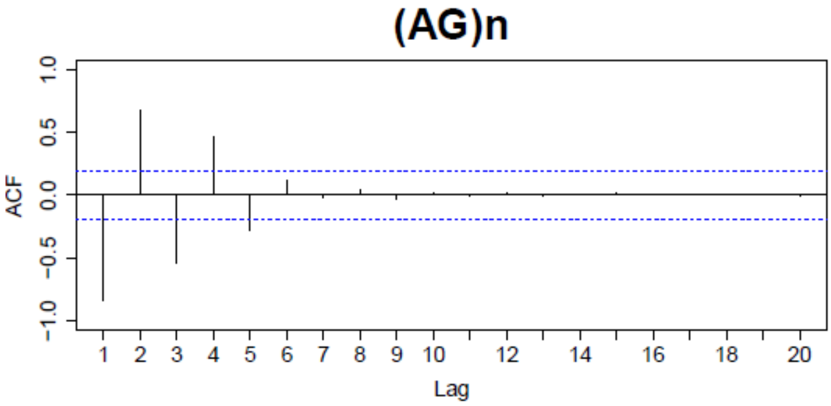


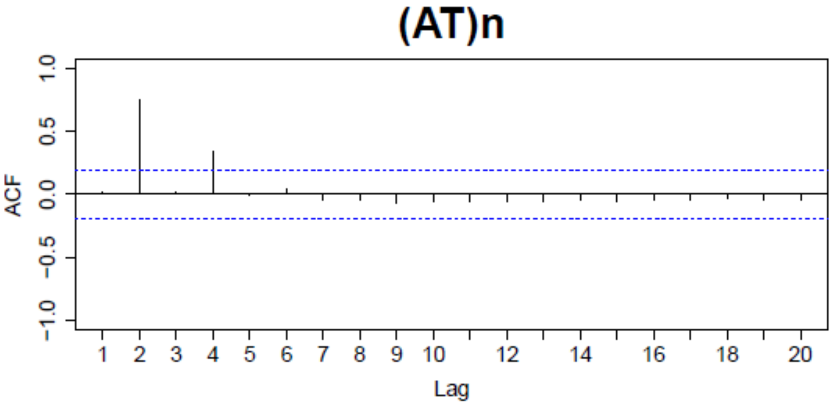
In summary we find that that, while read termination does not appear to explain the drop in read depth in G4 motifs, sequencing quality might contribute to it. Additionally, we reported in the manuscript an enrichment in deletions in reads mapping to G4 motifs, which can also contribute to lower read depth.

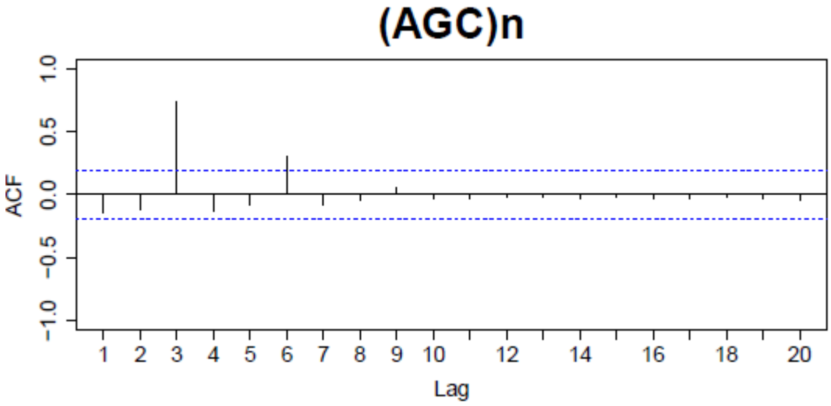
SUPPLEMENTARY NOTE 2

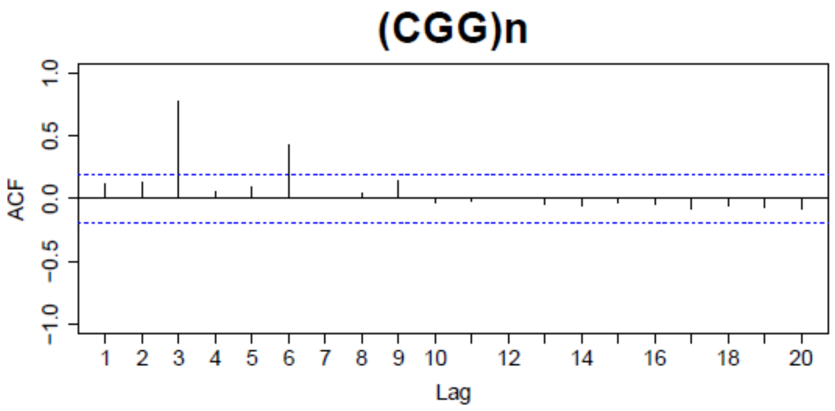
Periodicity in polymerization kinetics for microsatellites

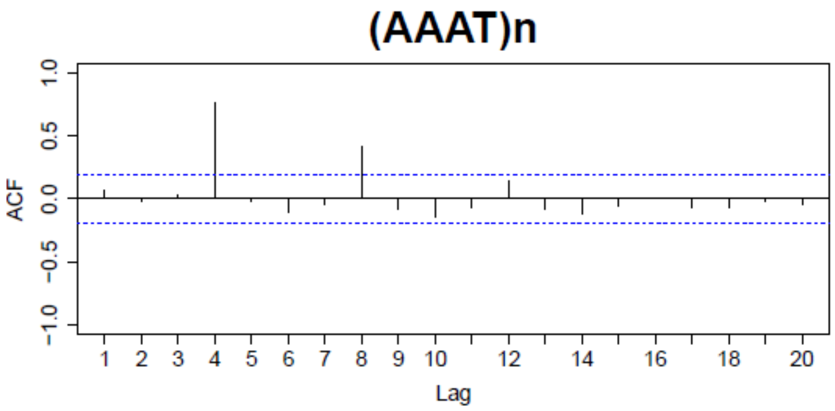
For many STRs the periodicity can be clearly seen in the autocorrelation plots: large autocorrelation at lags 2 and 4 for STRs of period 2 (i.e. periodic positive/negative signals); large autocorrelation at lags 3 and 6 for STRs of period 3; large autocorrelation at all lags for G4+ (suggesting a non-periodic trend in the median curve); small autocorrelation for motif-free regions (no strong trend or periodicity; an approximately constant median curve with ‘noise-like’ variation about it). Below are some examples of autocorrelation plots:

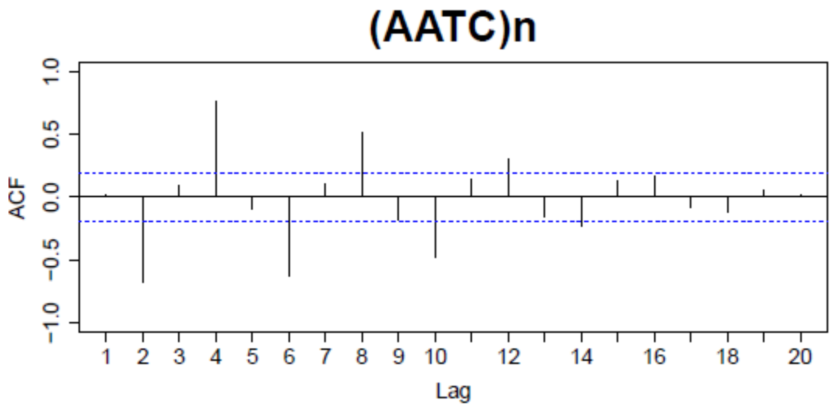


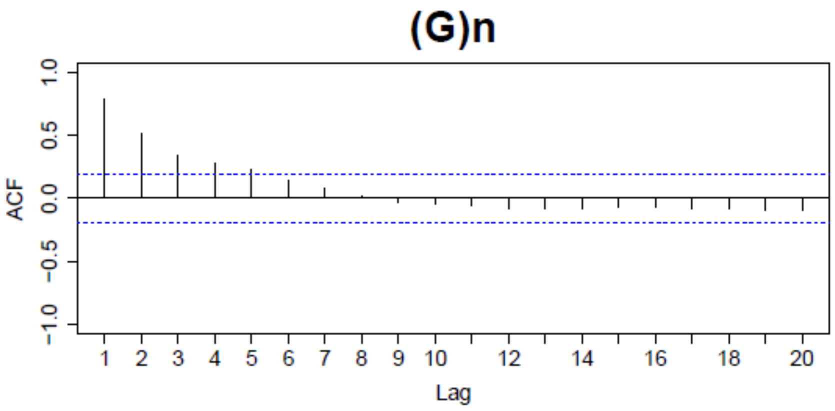


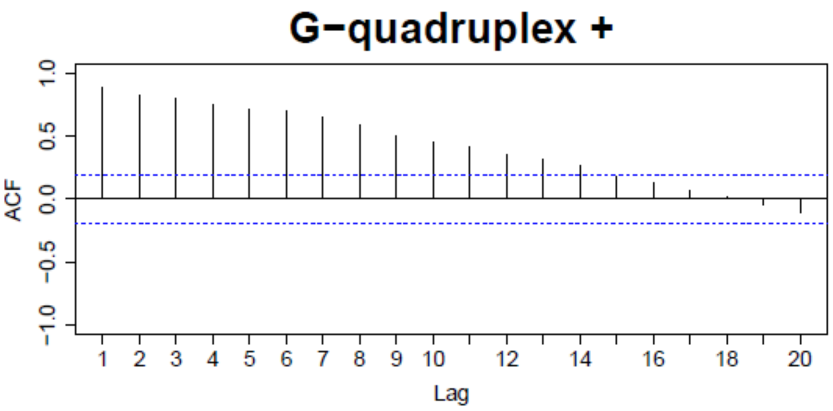


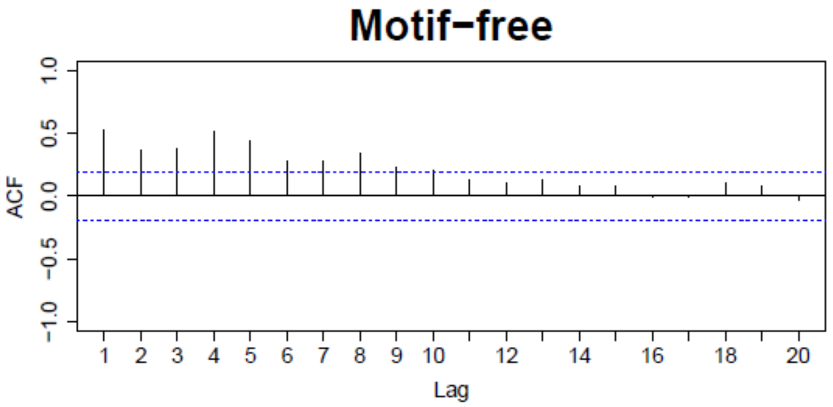

(CGG)n
ACF
1.0
0.5
0.0
−0.5
−1.0
1 2 3 4 5 6 7 8 9 10 12 14 16 18 20
Lag


(AAAT)n
ACF
1.0
0.5
0.0
−0.5
−1.0
1 2 3 4 5 6 7 8 9 10 12 14 16 18 20
Lag

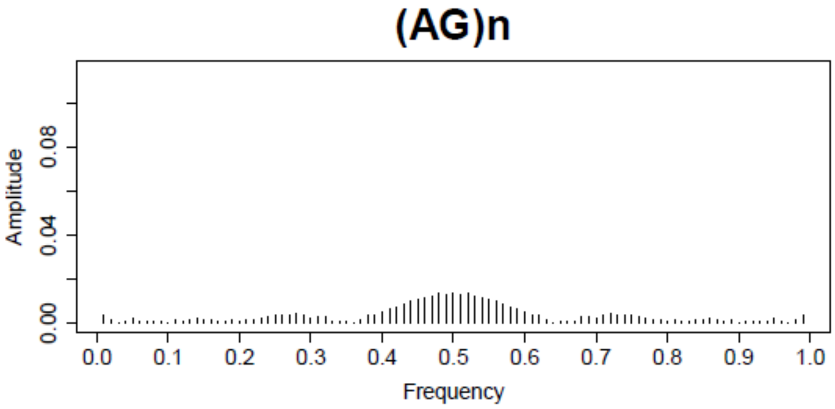

(AATC)n
ACF
1.0
0.5
0.0
−0.5
−1.0
1 2 3 4 5 6 7 8 9 10 12 14 16 18 20
Lag

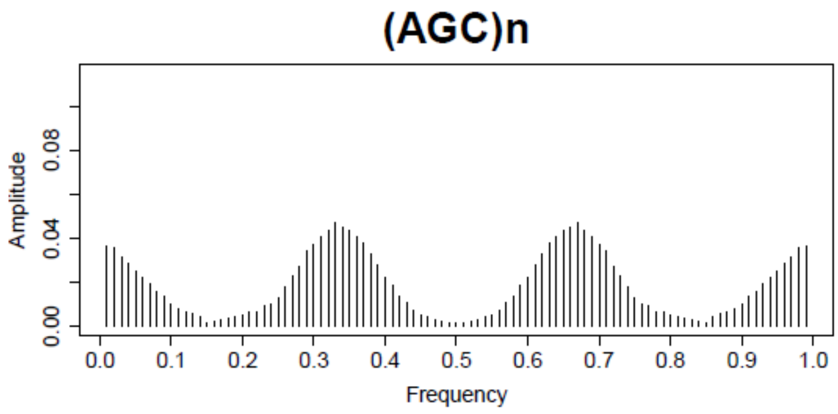

(G)n
ACF
1.0
0.5
0.0
−0.5
−1.0
1 2 3 4 5 6 7 8 9 10 12 14 16 18 20
Lag

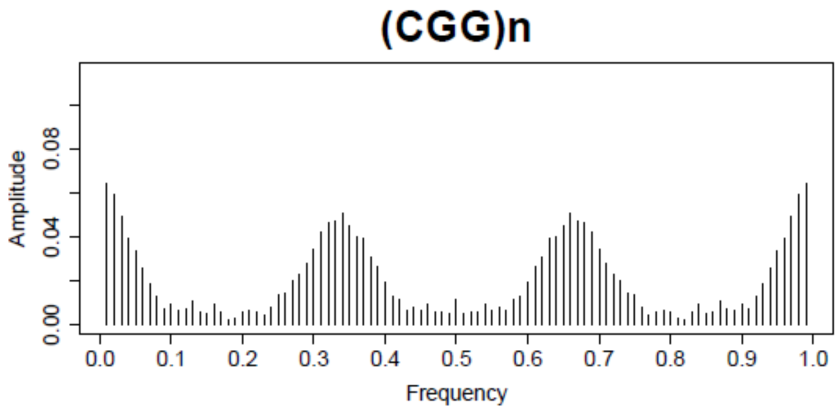


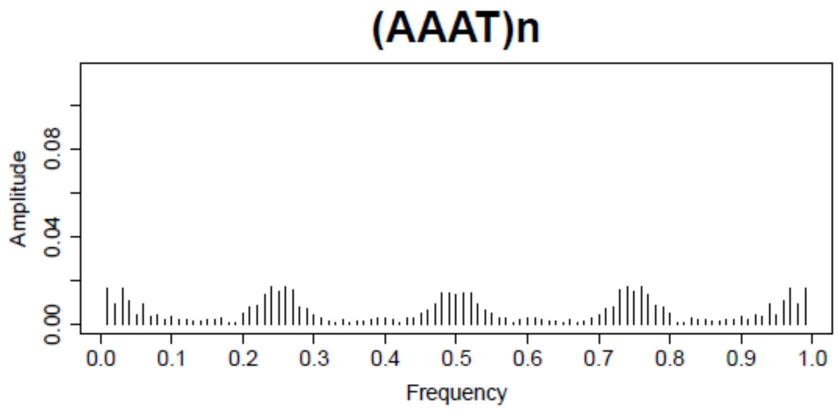


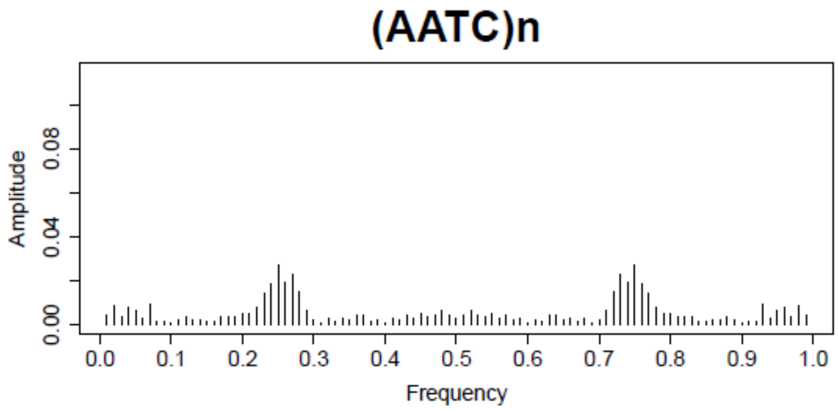
Discrete Fourier transforms also confirm periodicity in the amplitude vs frequency plots of STRs. Frequencies are computed with respect to 100-bp windows: 0.01 correspond to a period of 100 bp, 0.99 to a period of 1 bp, 0.5 to 2 bp, 0.33 to 3 bp and 0.66 to 1.5 bp. We observe large amplitude around frequency 0.5 for STRs of period 2; large amplitude around frequencies 0.33 and 0.66 for STRs of period 3; large amplitude at different frequencies close to 0 and 1 for G4+ (suggesting the presence of a non-periodic trend in the median curve); no large amplitude for motif-free regions (suggesting a constant median curve with 'noise-like' variation about it). Below are some examples of amplitude vs frequency plots:



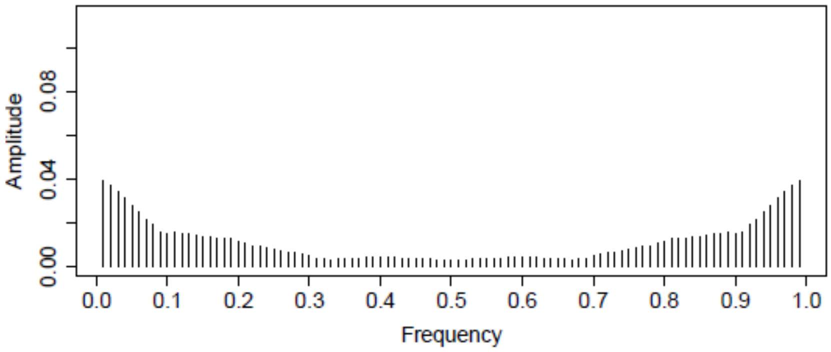

(AGC)n


(CGG)n

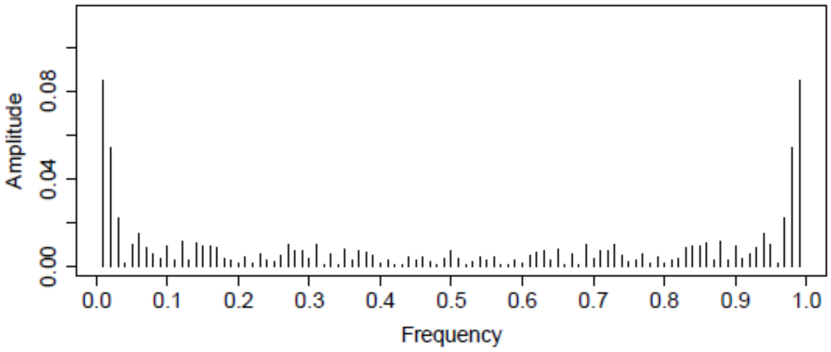

(AAAT)n


(AATC)n

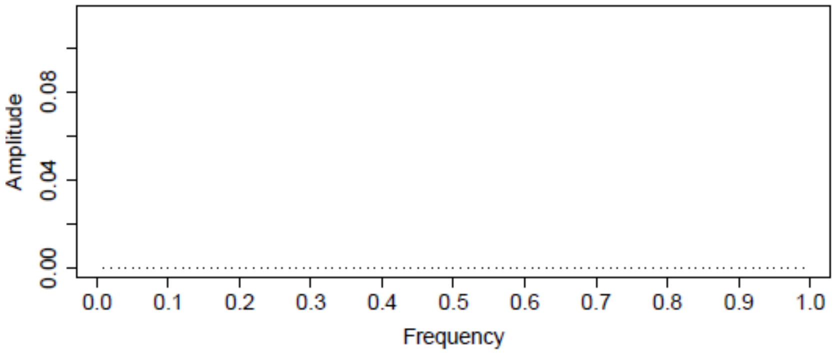
(G)n



G−quadruplex +



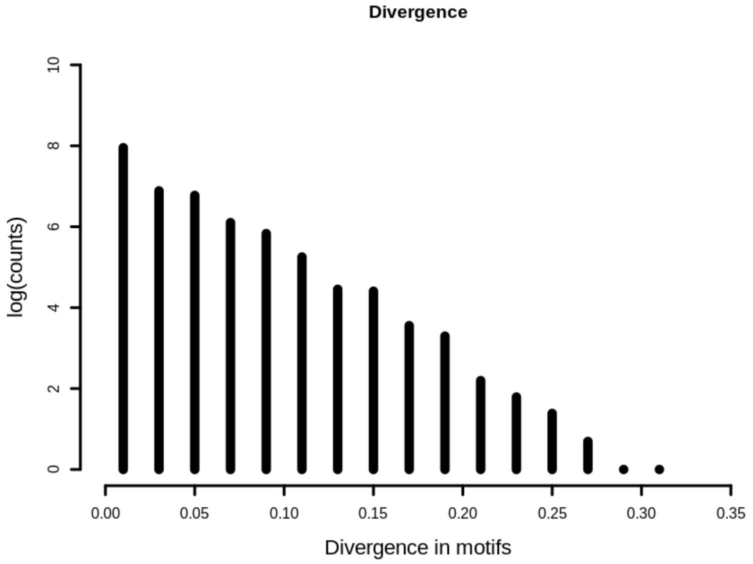
Motif−free

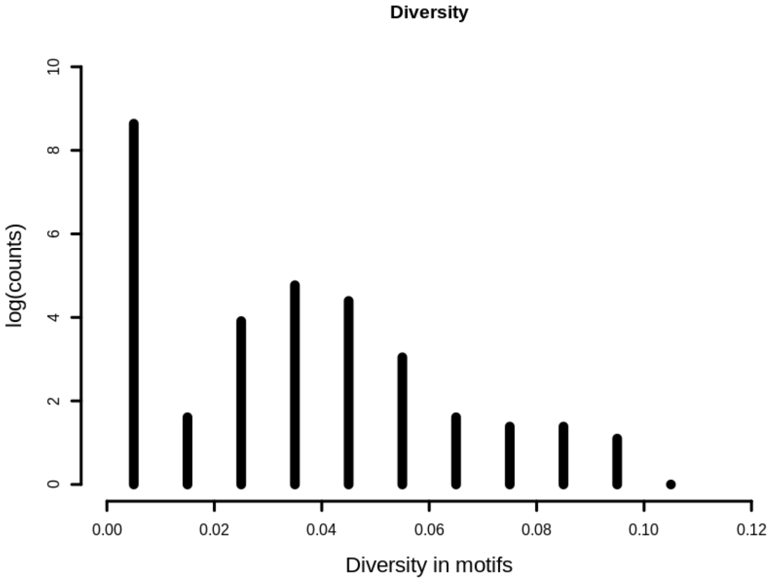


SUPPLEMENTARY NOTE 3

Contrasting extremes in of zero-inflated distributions of divergence and diversity

The plots below show that the distributions of divergence and diversity for G4 are strongly enriched in zeroes; many motifs do not harbor any variants (diversity and/or divergence = 0; note that the vertical axes of the plots, representing counts, are on the log scale). Thus, analyses based on the entire distributions would be heavily affected by invariant motifs; this is why we focused on contrasting the extremes of the distributions.





We also investigated whether the use of different thresholds in defining the extreme ranges would lead to different results; see Table below.

Median p-value of 1000 bootstrap *t*-test between top and bottom 25%, 7%, 5%, 4%, 3%, and 2%

		Divergence	Diversity
Error rate	25%	0.035	0.421
	7%	0.029	0.416
	5%	0.042	0.421
	4%	0.194	6.502e-04
	3%	**0.035**	**2.153e-04**
	2%	0.069	1.833e-05
Residuals IPD	25%	2.997e-06	0.591
	7%	1.284e-04	0.587
	5%	3.816e-04	0.590
	4%	1.869e-04	0.011
	3%	**2.228e-05**	**0.021**
	2%	0.001	0.008

In these comparisons, to match the number of motifs in top and bottom extreme groups, we subsampled motifs with divergence or diversity equal to zero -- and repeated a bootstrap *t*-test contrasting each of 1,000 bottom subsamples against the top group. The Table reports median p-values from 1,000 tests. The divergence results remain significant at all percentage cutoffs used to define top and bottom groups except at 4% for error rate (they are marginally significant at 2%). The diversity results fare significant for cutoffs of 4% or less, but not for larger cutoffs. This is because as many as 95% of the motifs have diversity = 0: as the percent cutoff to define top and bottom groups grows, the two become harder to separate both in terms of error rate and in terms of residuals IPD. We now results for the comparison of the top and bottom groups defined with a 3% cutoff -- both for divergence and for diversity. We feel that this is a reasonable choice -- which allow us to include two lines of evidence based on two independent data sets (the one for divergence and the one for diversity), which are in agreement. Taking higher cutoffs, in the case of diversity, breaks this agreement only because we do not have enough variant-containing motifs to place in the top group.

SUPPLEMENTARY NOTE 4

Analysis of the potential contribution of increased error rate at G4 motifs to the observed variants in the 1,000 Genomes Project

The 1000 Genomes Project employs Illumina technology, that is much more accurate than SMRT, having a single pass error rate of about 0.1%. In this work we restricted our attention to SMRT sequencing, but it is possible that non-B DNA also affects error rates of other sequencing technologies using different polymerases. As a consequence, the impact of Illumina errors on the detection of variants might be higher in non-B DNA motifs than in motif-free regions.

Indeed, a previous study of HiSeq data (Schirmer et al. 2016) demonstrated that the occurrence of Illumina errors depends on sequence context. In particular, substitutions were shown to depend on the 3-mers preceding them. The datasets analyzed in this 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 the impact of this bias on high-frequency variant calling (global frequency≥0.05) in the 1000 Genomes Project, we considered the following binomial model for sequencing errors. We modeled the Illumina sequencing error process at a nucleotide belonging to a G-quadruplex as a Bernoulli trial $X_{G4} \sim B(1, f_{G4})$, with f_{G4} representing the error rate per read base. Similarly, we considered 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 $X_{G4} \sim B(1, f_{G4})$ and $X_C \sim B(1, f_C)$, for nucleotides belonging to G-quadruplexes 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 $V_{G4} \sim B(5{,}008; f_{G4})$ for G-quadruplexes 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 (global frequency≥0.05) is due solely to sequencing error can be computed as $P(V_{G4} \geq 251)$ and $P(V_C \geq 251)$, where 251 = 5,008×0.05 .

We assume, again as a worst case scenario, that all 322 of the 1000 Genome Project variants observed in G-quadruplexes (G4+) with global frequency >0.05 occurred in nucleotides immediately following an occurrence of the "GGG" 3-mer. In this positions the Illumina error rate can be as high as 0.035, and the expected number of variants due to sequencing errors is equal to 7.9×10^{-6}. Although the number of expected variants is higher following this 3-mer, it is still very low (less than 1 observed variant is expected because of errors).

The conservative calculations presented here show that high-frequency substitutions detected within G-quadruplexes based on the 1000 Genome Project data are extremely unlikely to be caused by systematic sequence context biases in Illumina sequencing errors.

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 (Cer et al. 2013) (and according to STR-FM (Fungtammasan et al. 2015) 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 (Sinden 2012); a cruciform)	$(A)_n$ (Ref (Sinden 2012); also form A tract/bent)	$(AC)_n$ (Ref (Sinden 2012))
$(AAT)_n$ (predicted from sequence)	$(C)_n$ (Ref (Sinden 2012))	$(CG)_n$ (Ref (Sinden 2012))
$(ACT)_n$ (predicted from sequence)	$(G)_n$ (Ref (Sinden 2012); also form A tract/bent)	$(GT)_n$ (Ref (Sinden 2012))
$(AGC)_n$ (Ref (Mirkin and Mirkin 2007))	$(T)_n$ (Ref (Sinden 2012))	
$(AGG)_n$ (Ref (Huertas and Azorín 1996))	$(AG)_n$ (Ref (Sinden 2012))	
$(AGT)_n$ (predicted from sequence)	$(CT)_n$ (Ref (Sinden 2012))	
$(ATC)_n$ (predicted from sequence)	$(AAG)_n$ (Ref (Sinden 2012))	
$(ATG)_n$ (predicted from sequence)	$(CCT)_n$ (predicted from sequence)	
$(ATT)_n$ (Ref (Trotta et al. 2000))	$(CTT)_n$ (Ref (Sinden 2012))	
$(CCG)_n$ (Ref (Mirkin and Mirkin 2007))		
$(CGG)_n$ (Ref (Mirkin and Mirkin 2007))		
$(CTG)_n$ (Ref (Mirkin and Mirkin 2007))		

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 GGGAGGGAGGTGGGGGGGG	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
$(GGT)_5$ GGTGGTGGTGGTGGT	45.2	bimol	138	paral
$(GGT)_6$ GGTGGTGGTGGTGGTGGT	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 SMRT sequencing errors.

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

Table S8. 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`

Table S9. Kinetics and error rates in G-quadruplexes are linked to their divergence and diversity. Residual IPDs were obtained as differences between observed mean IPDs and the ones predicted according to mononucleotide sequence composition (see Methods). Highly diverged/diverse G-quadruplexes (top 3%) have higher IPD and SMRT mismatch error rates than the ones with low divergence/diversity (bottom 3%). Sample sizes: n=314 (in the top 3% of divergence; the same number of motifs in the bottom 3% of divergence); n=302 (in the top 3% of diversity; the same number of motifs in the bottom 3% of diversity). The choice of 3% most extreme was driven by the shape of the distributions of Divergence and Diversity. For instance, the trend and statistical significance in Divergence remains using the 25% most extremes, while it disappears in Diversity since the 95th percentile (5% highest) is 0.

	Divergence		**Diversity**	
	Bottom 3%	**Top 3%**	**Bottom 3%**	**Top 3%**
Residuals log mean IPDs	0.1692	0.2771	0.1732	0.249
***t*-test p-value**	$2.228\text{x}10^{-04}$		0.021	
Log error rates	0.0226	0.0247	0.0223	0.0252
***t*-test p-value**	0.035		$2.153\text{x}10^{-04}$	

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.

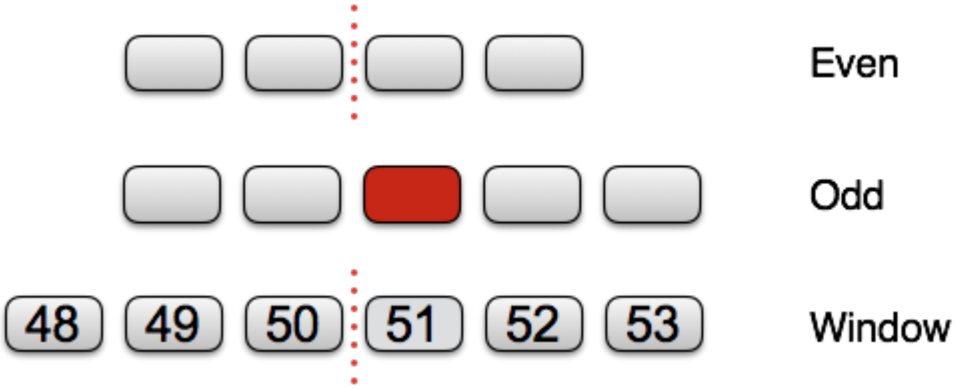
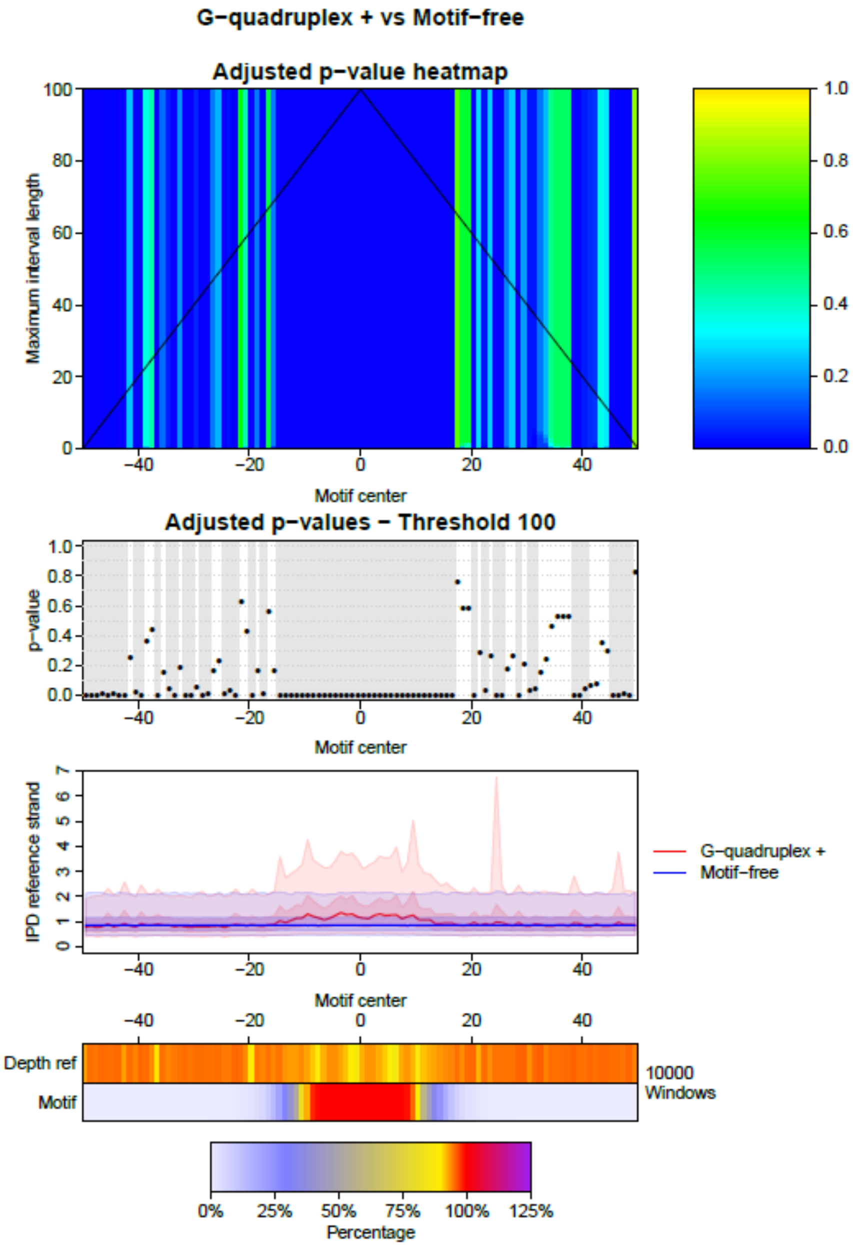


Figure S2. An example of detailed results of Interval-Wise Testing.
Results of IWT using multi-quantile statistic and a random subsample of 10,000 windows for the comparisons **A** G-quadruplex motifs on reference strand vs. motif-free windows. **B** $(AGC)_n$ vs. motif-free

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

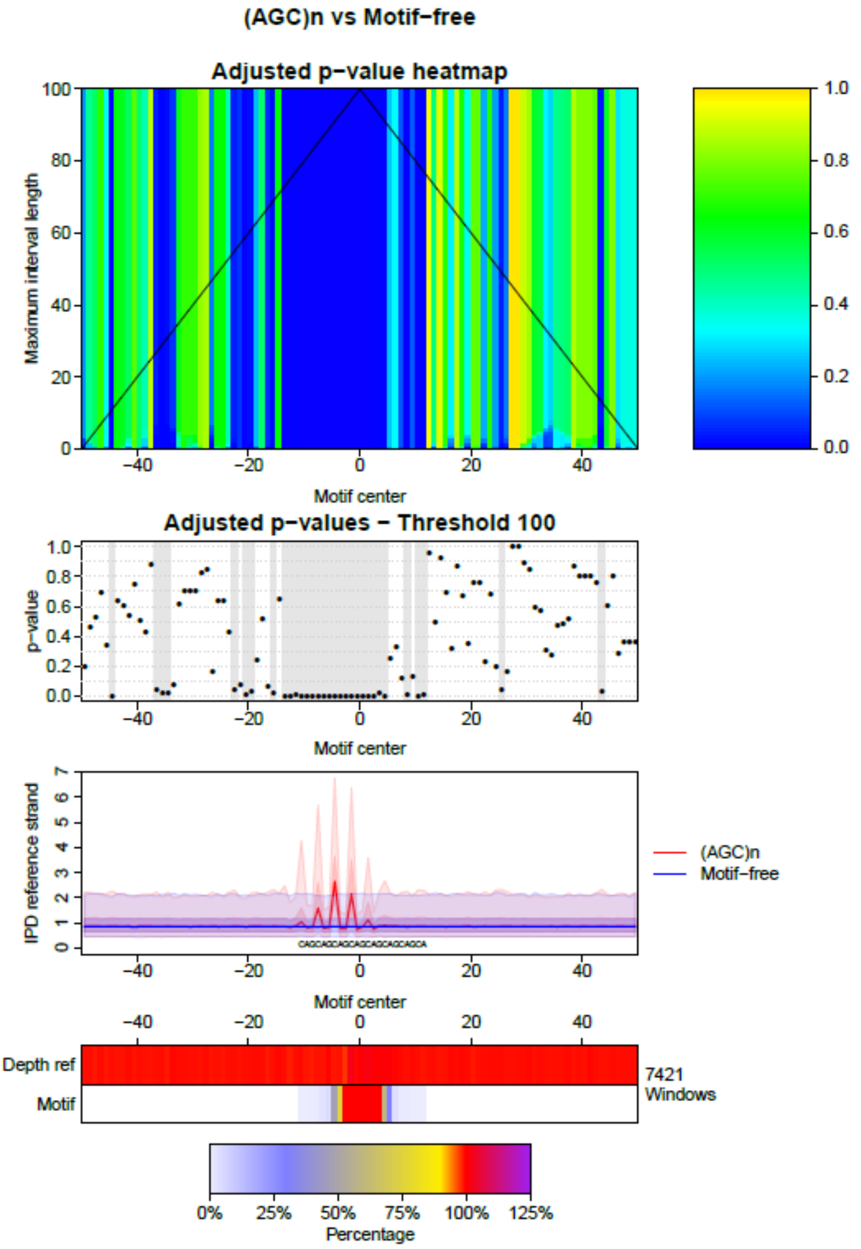
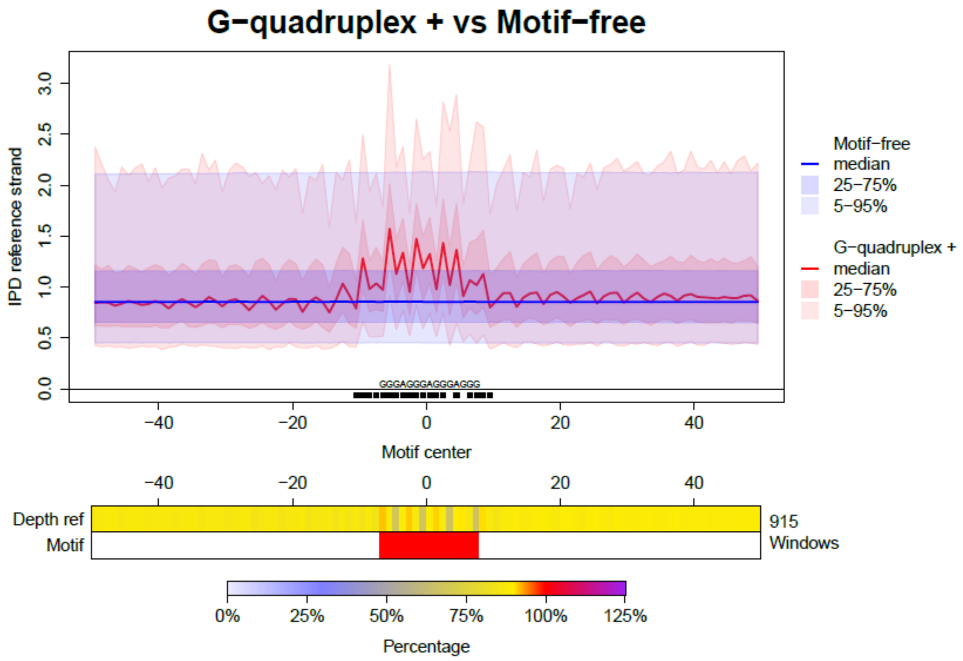
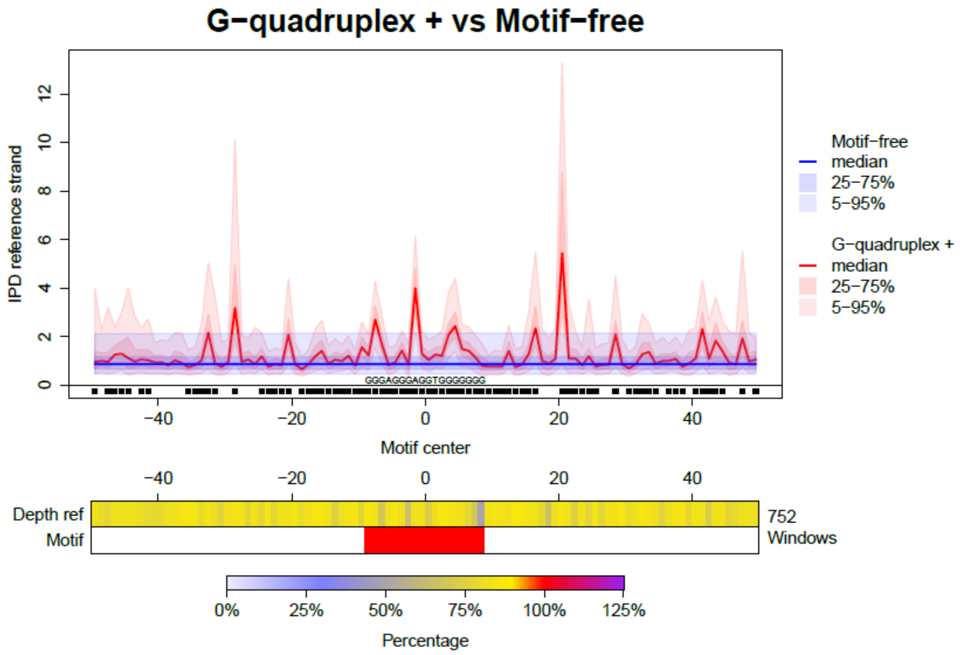

(AGC)n vs Motif−free
Adjusted p−value heatmap
Maximum interval length
0
20
40
60
80
100
−40
−20
0
20
40
Motif center
0.0
0.2
0.4
0.6
0.8
1.0
Adjusted p−values − Threshold 100
p−value
0.0
0.2
0.4
0.6
0.8
1.0
−40
−20
0
20
40
Motif center
IPD reference strand
0
1
2
3
4
5
6
7
CAGCAGCAGCAGCAGCAGCAGCAGCA
−40
−20
0
20
40
Motif center
(AGC)n
Motif−free
−40
−20
0
20
40
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$_{3\text{-}5}$G$_3$ motifs only have the central elevation and lack the 3’ spike. **B** GGGA$_2$GGT$_1$G$_{7\text{-}8}$ and **C** G$_3$T$_1$G$_2$A$_1$G$_3$T$_1$G$_3$A$_1$G$_2$A$_1$G$_3$ present only spikes in 5’, 3’ and overlapping the motif. **D** G$_4$TN$_1$G$_5$A$_1$G$_6$A$_1$G$_3$, **E** G$_4$T$_1$G$_5$A$_2$G$_6$A$_1$G$_3$, **F** G$_4$T$_1$G$_6$A$_{1\text{-}2}$G$_5$A$_1$G$_3$, **G** G$_4$T$_1$G$_6$AGN$_1$G$_4$A$_1$G$_3$, **H** G$_4$T$_1$G$_6$A$_1$G$_5$A$_{1\text{-}2}$G$_3$ and **I** G$_4$T$_1$G$_6$AT$_1$G$_5$A$_1$G$_3$ all have a central elevation surrounded by spikes as well as the 3’ spike. Finally, **J** GGGT$_3$GGG$_1$ 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.

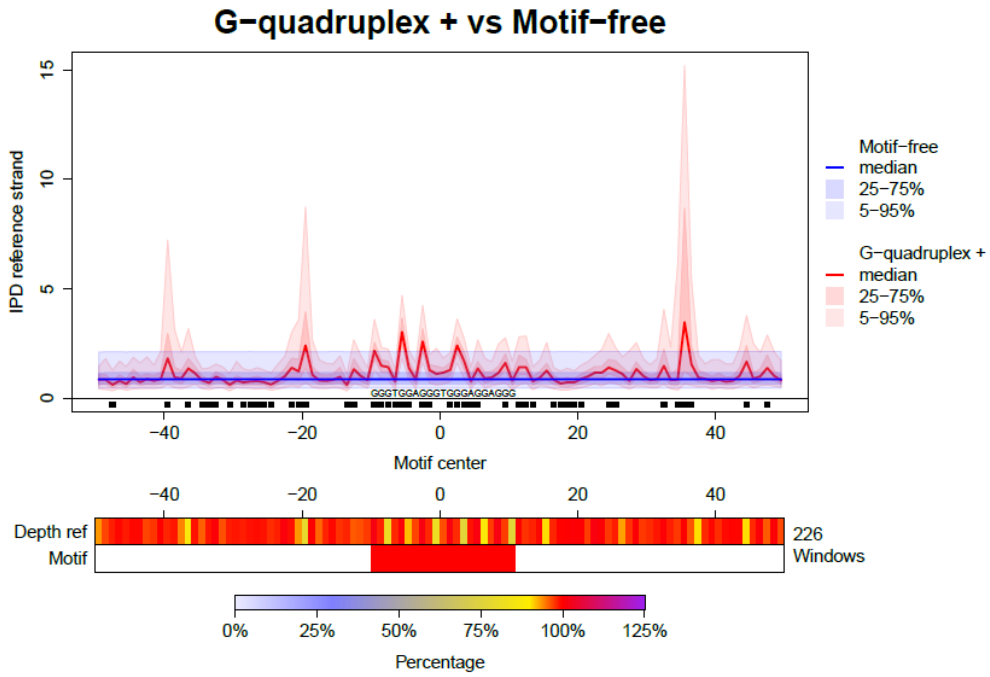
A



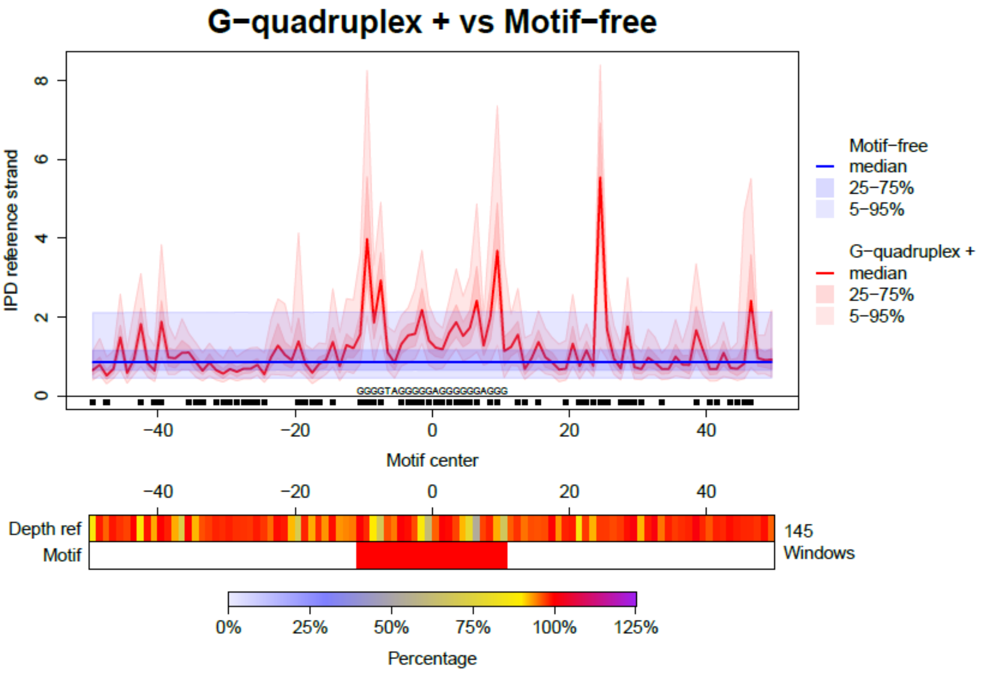
B


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

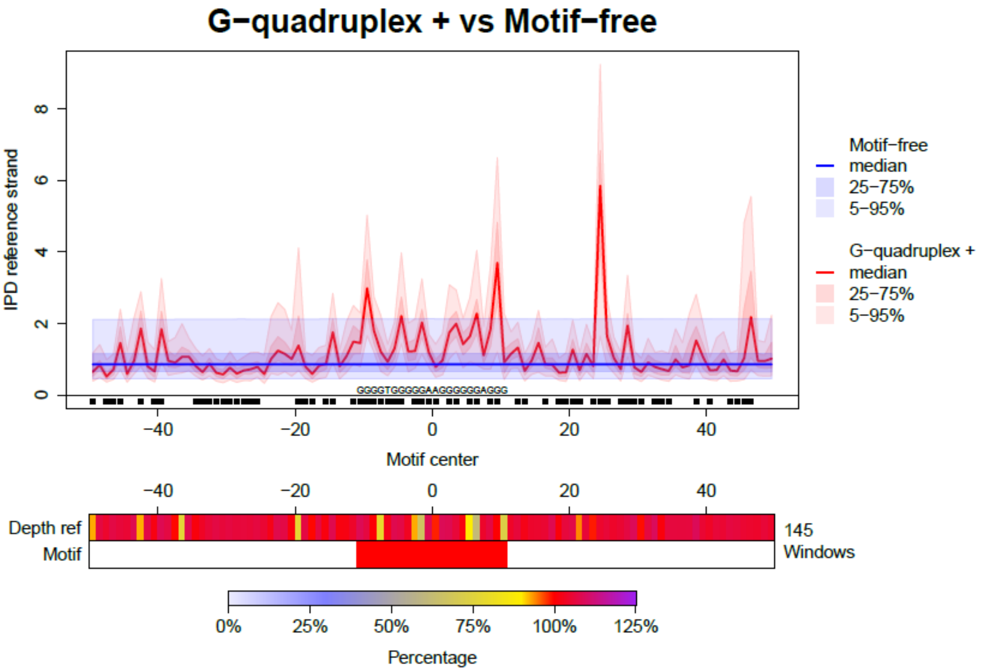
C


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

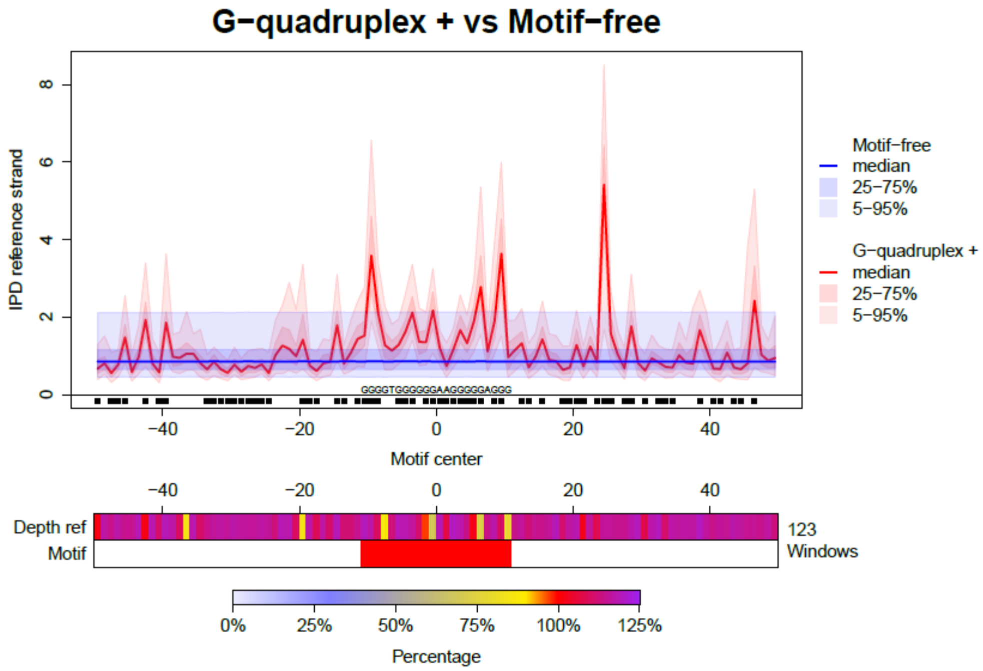
D


G−quadruplex + vs Motif−free
IPD reference strand
0
2
4
6
8
GGGGTAGGGGGAGGGGGGAGGG
−40
−20
0
20
40
Motif center
Motif−free
median
25−75%
5−95%
G−quadruplex +
median
25−75%
5−95%
−40
−20
0
20
40
Depth ref
Motif
145
Windows
0%
25%
50%
75%
100%
125%
Percentage

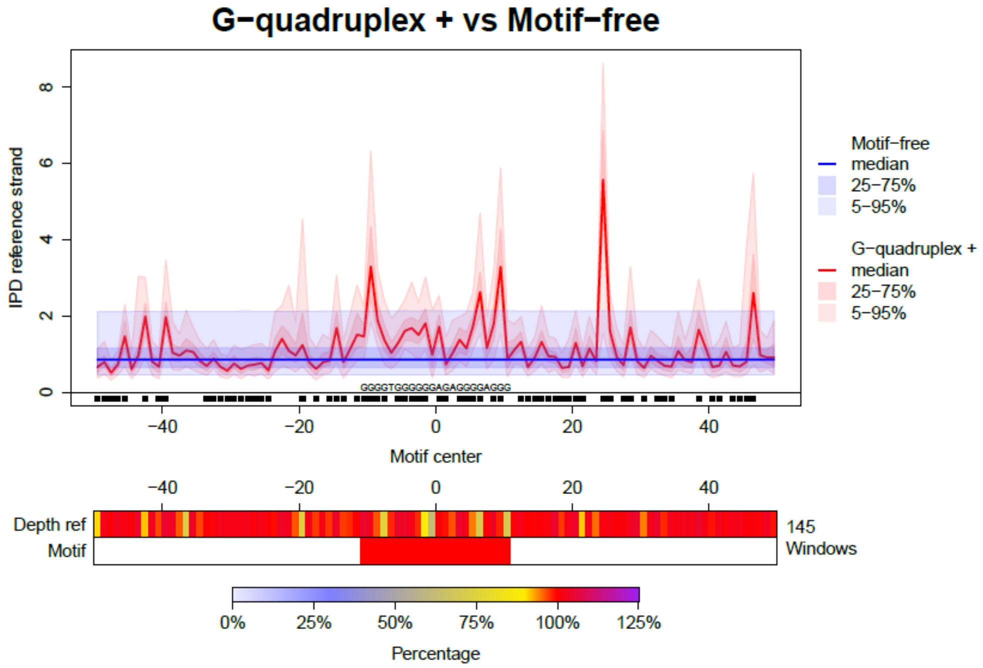
E


G−quadruplex + vs Motif−free
IPD reference strand
0
2
4
6
8
GGGGTGGGGGAAGGGGGGAGGG
−40
−20
0
20
40
Motif center
Motif−free
median
25−75%
5−95%
G−quadruplex +
median
25−75%
5−95%
−40
−20
0
20
40
Depth ref
Motif
145
Windows
0%
25%
50%
75%
100%
125%
Percentage

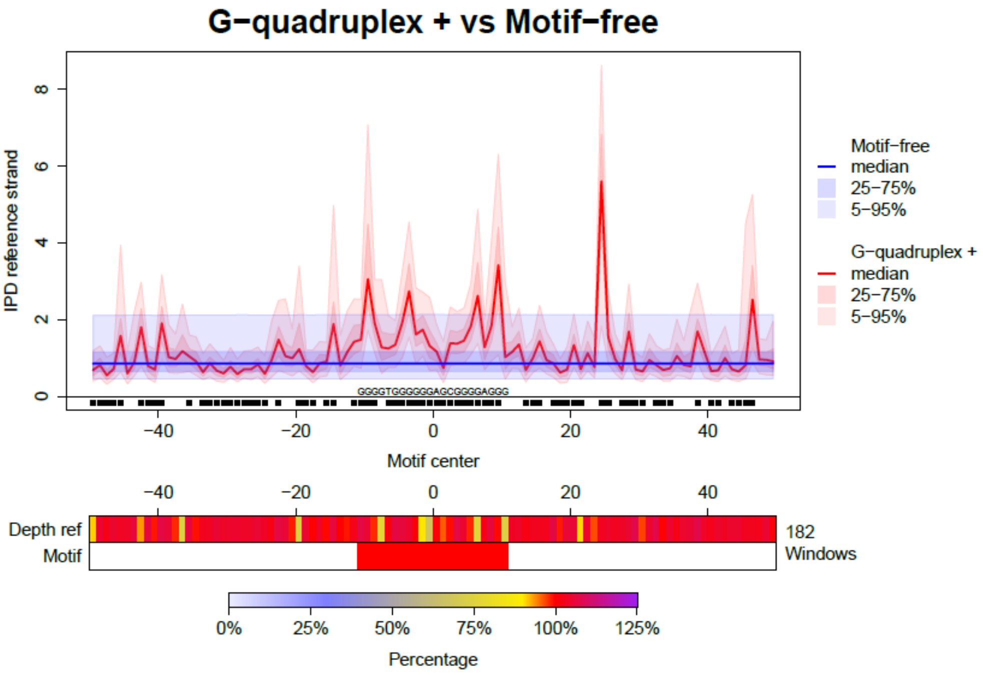
F


G−quadruplex + vs Motif−free
IPD reference strand
0
2
4
6
8
GGGGTGGGGGGAAGGGGGAGGG
−40
−20
0
20
40
Motif center
Motif−free
median
25−75%
5−95%
G−quadruplex +
median
25−75%
5−95%
−40
−20
0
20
40
Depth ref
Motif
123
Windows
0%
25%
50%
75%
100%
125%
Percentage

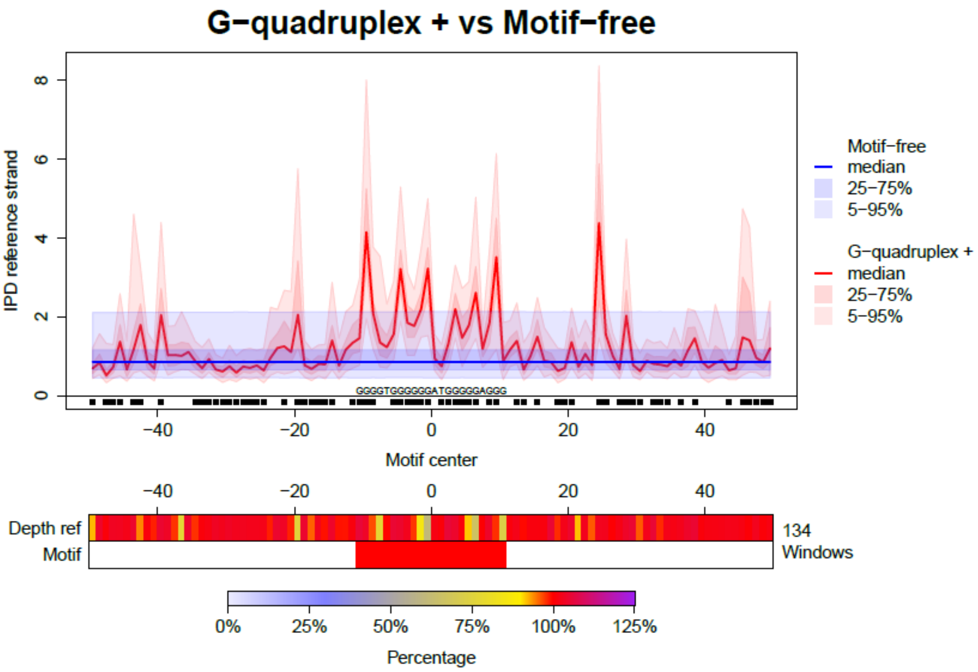
G


G−quadruplex + vs Motif−free
IPD reference strand
0
2
4
6
8
GGGGTGGGGGGAGAGGGGAGGG
−40
−20
0
20
40
Motif center
Motif−free
median
25−75%
5−95%
G−quadruplex +
median
25−75%
5−95%
−40
−20
0
20
40
Depth ref
Motif
145
Windows
0%
25%
50%
75%
100%
125%
Percentage

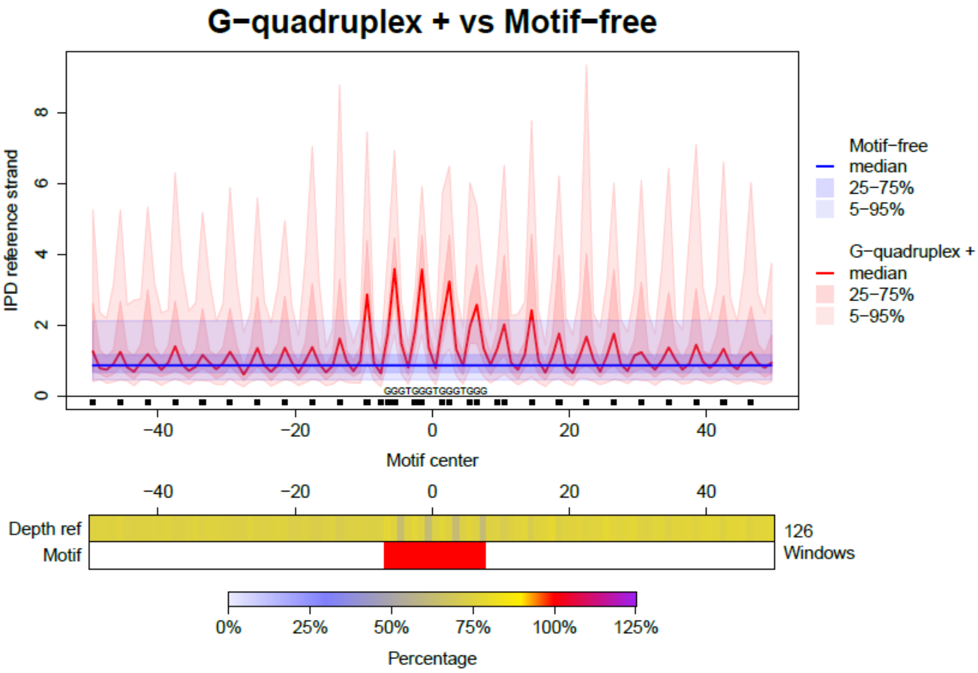
H


G−quadruplex + vs Motif−free
IPD reference strand
0
2
4
6
8
GGGGTGGGGGGAGCGGGGAGGG
−40
−20
0
20
40
Motif center
Motif−free
median
25−75%
5−95%
G−quadruplex +
median
25−75%
5−95%
−40
−20
0
20
40
Depth ref
Motif
182
Windows
0%
25%
50%
75%
100%
125%
Percentage

I


G−quadruplex + vs Motif−free
IPD reference strand
0
2
4
6
8
GGGGTGGGGGGATGGGGGAGGG
−40
−20
0
20
40
Motif center
Motif−free
median
25−75%
5−95%
G−quadruplex +
median
25−75%
5−95%
−40
−20
0
20
40
Depth ref
Motif
134
Windows
0%
25%
50%
75%
100%
125%
Percentage

J

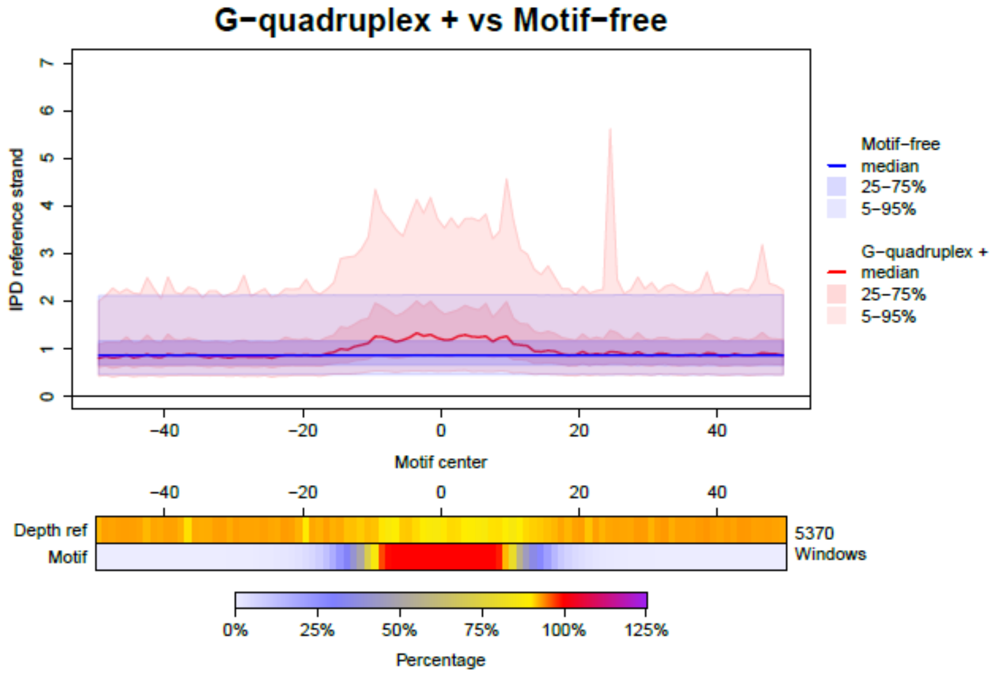


G−quadruplex + vs Motif−free
IPD reference strand
0
2
4
6
8
GGGTGGGTGGGTGGG
−40
−20
0
20
40
Motif center
Motif−free
median
25−75%
5−95%
G−quadruplex +
median
25−75%
5−95%
−40
−20
0
20
40
Depth ref
Motif
126
Windows
0%
25%
50%
75%
100%
125%
Percentage

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



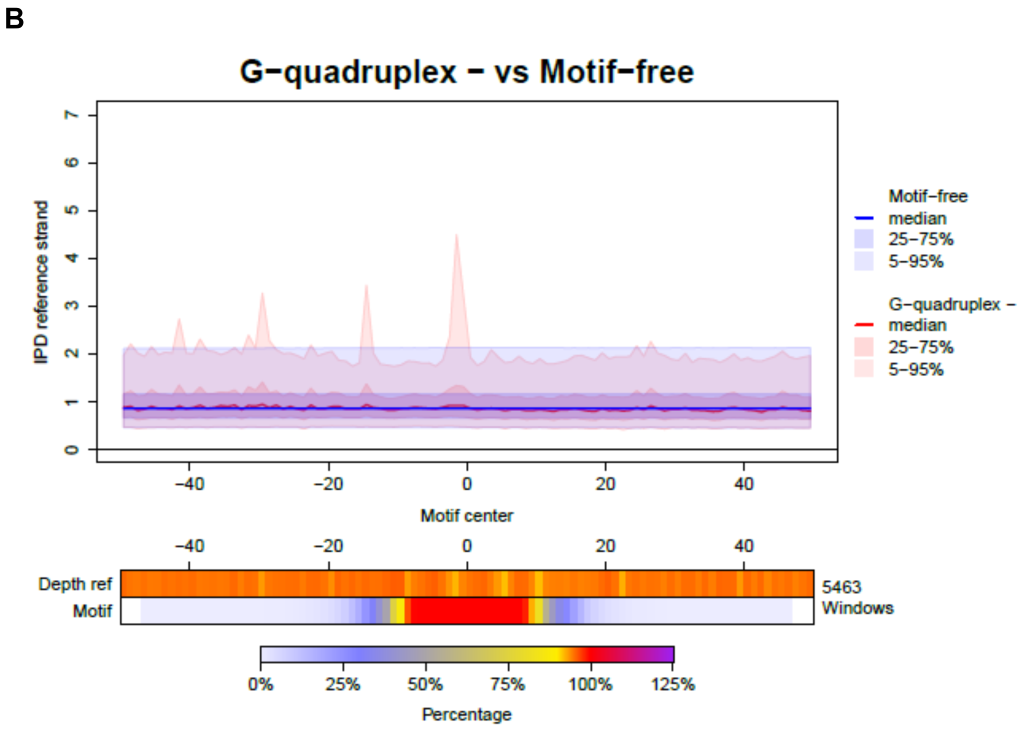
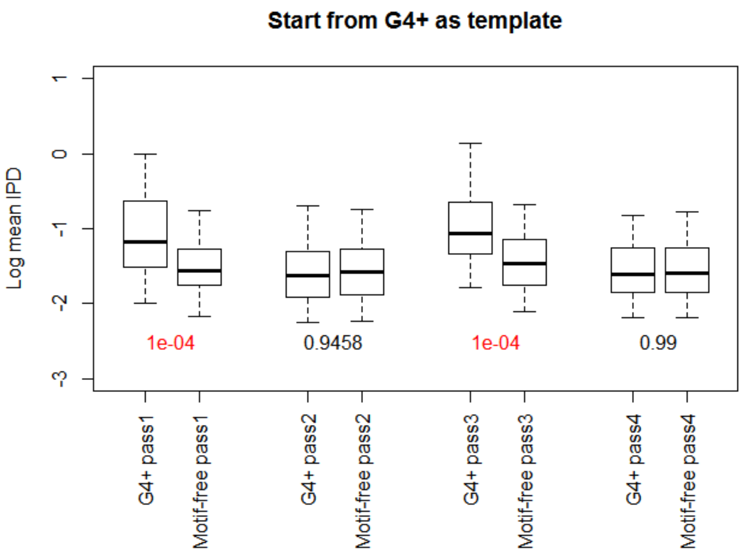

B
G−quadruplex − vs Motif−free
IPD reference strand
0
1
2
3
4
5
6
7
−40
−20
0
20
40
Motif center
Motif−free
median
25−75%
5−95%
G−quadruplex −
median
25−75%
5−95%
−40
−20
0
20
40
Depth ref
Motif
5463
Windows
0%
25%
50%
75%
100%
125%
Percentage

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



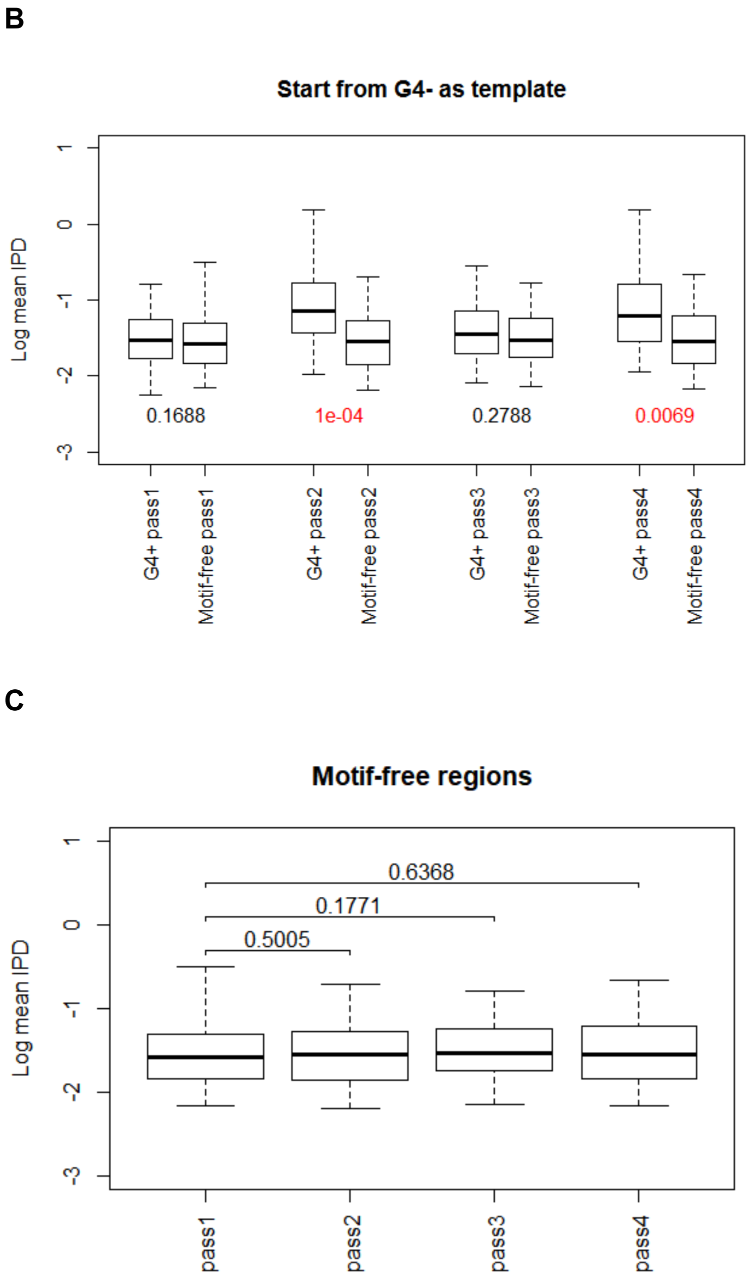
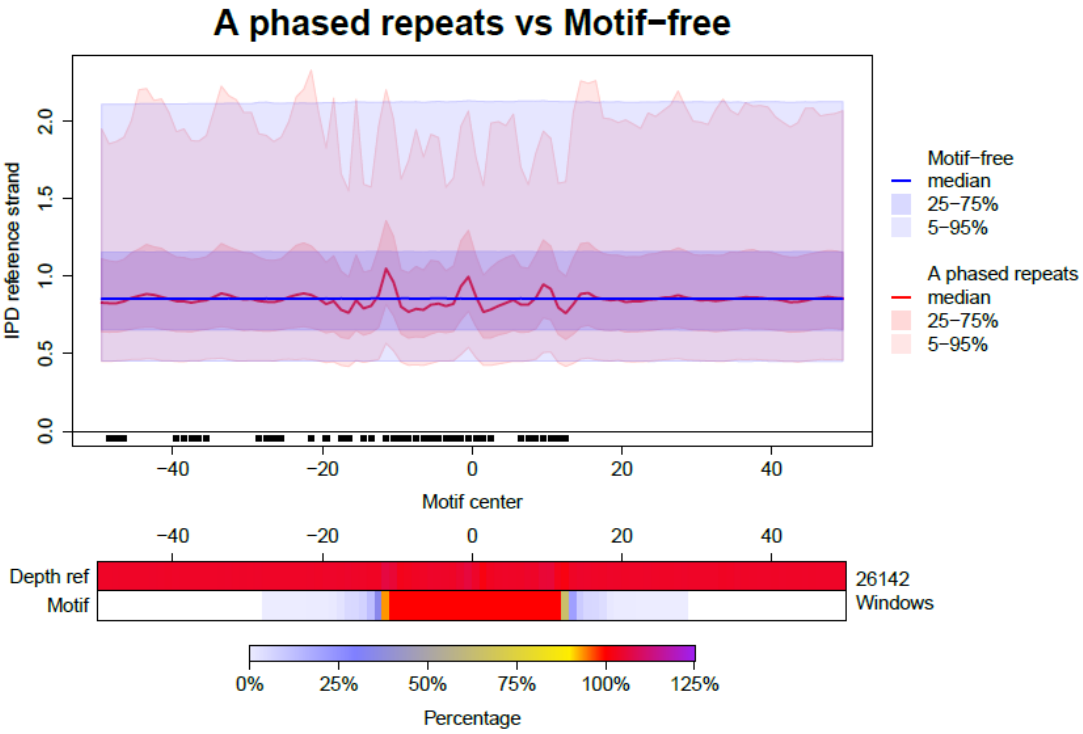

B
Start from G4- as template
Log mean IPD
1
0
-1
-2
-3
0.1688
1e-04
0.2788
0.0069
G4+ pass1
Motif-free pass1
G4+ pass2
Motif-free pass2
G4+ pass3
Motif-free pass3
G4+ pass4
Motif-free pass4
C
Motif-free regions
Log mean IPD
1
0
-1
-2
-3
0.6368
0.1771
0.5005
pass1
pass2
pass3
pass4

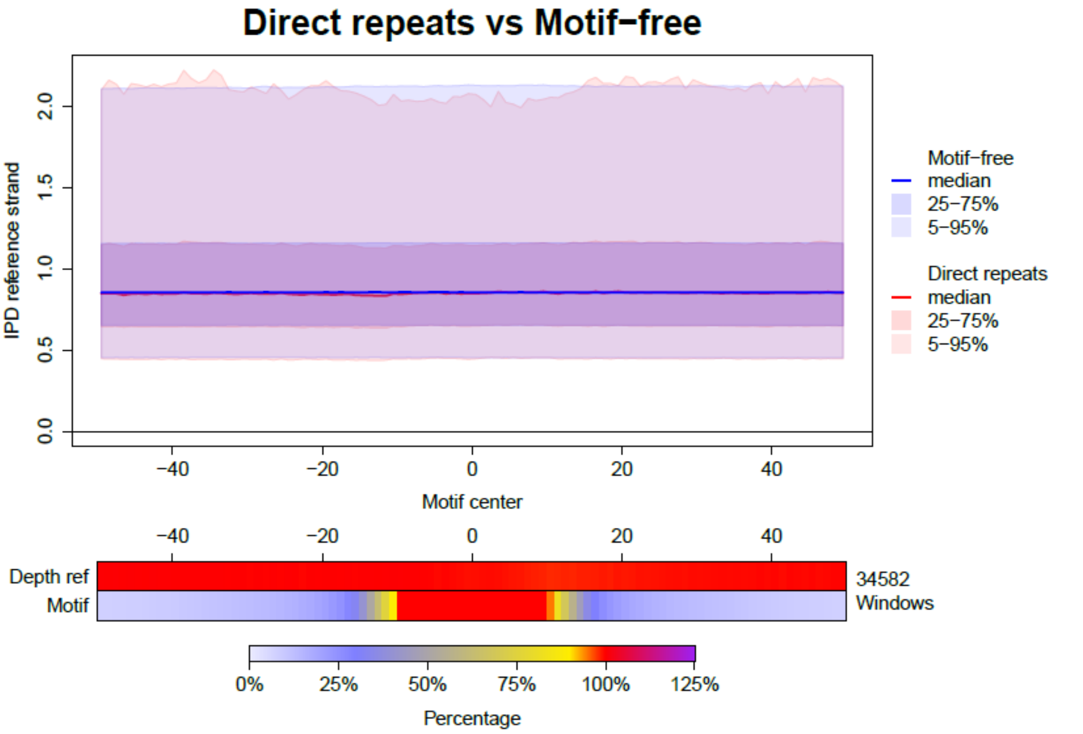
Figure S6. Effect of different non-B DNA motifs on IPDs.

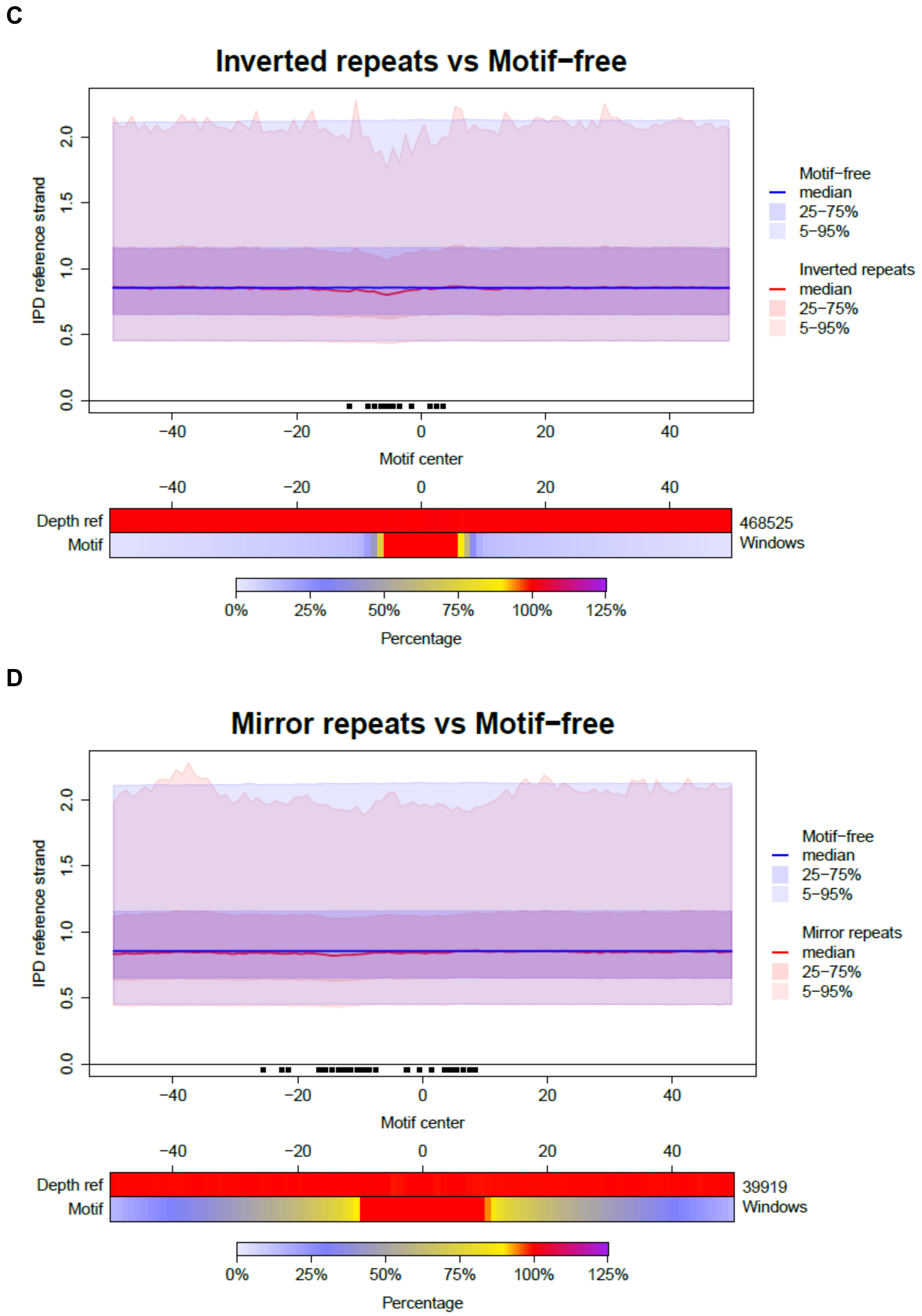
A A-phased repeats depress the IPD distribution. **B** Direct Repeats do not significantly change the IPD distribution. **C** Inverted Repeats depress the IPD distribution slightly. **D** Mirror Repeats slightly depress the IPD distribution. **E** Z-DNA motifs slightly increase the IPD distribution in 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
Motif−free
median
25−75%
5−95%
Inverted repeats
median
25−75%
5−95%
Depth ref
Motif
468525
Windows
0%
25%
50%
75%
100%
125%
Percentage
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
Motif−free
median
25−75%
5−95%
Mirror repeats
median
25−75%
5−95%
Depth ref
Motif
39919
Windows
0%
25%
50%
75%
100%
125%
Percentage

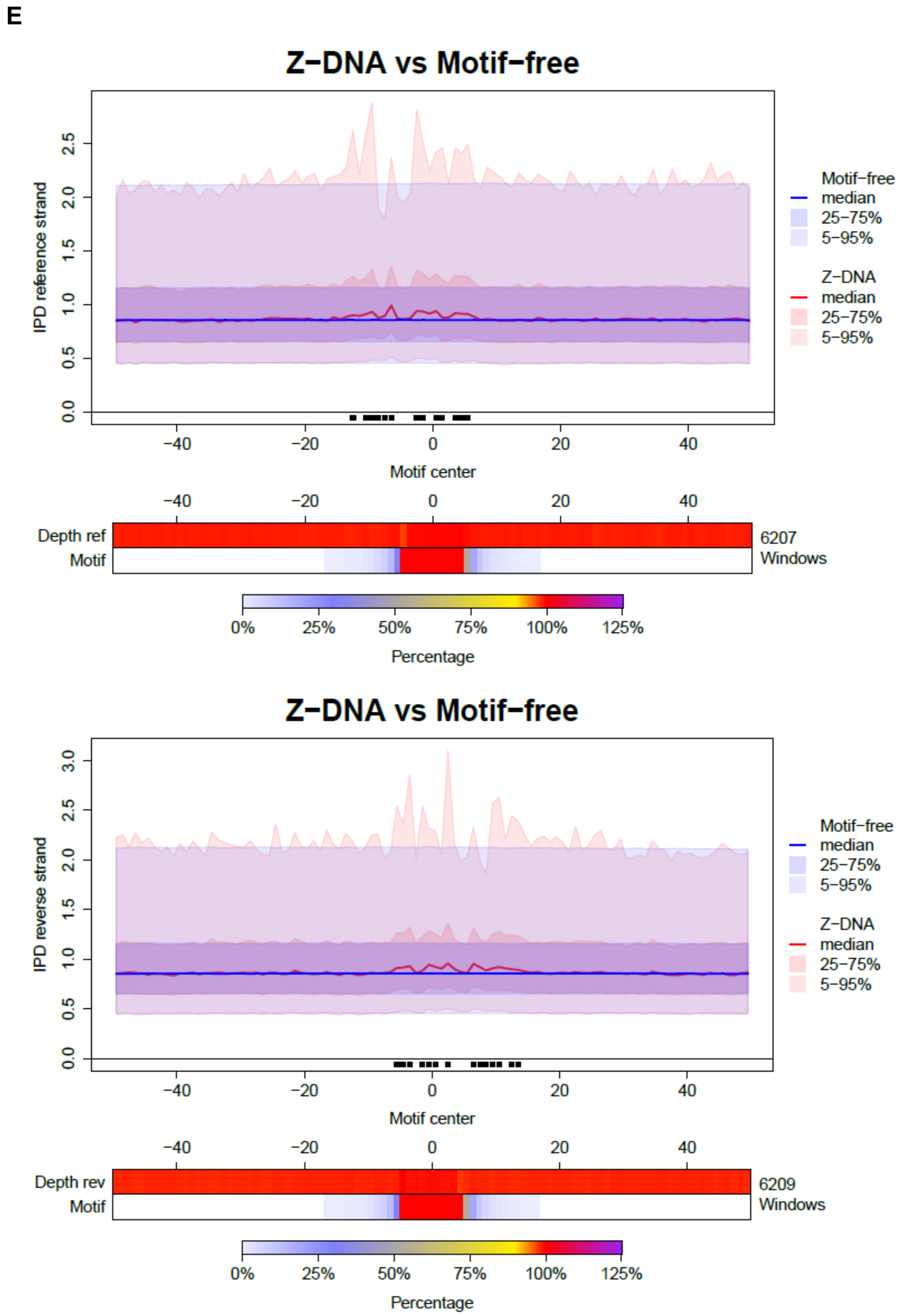
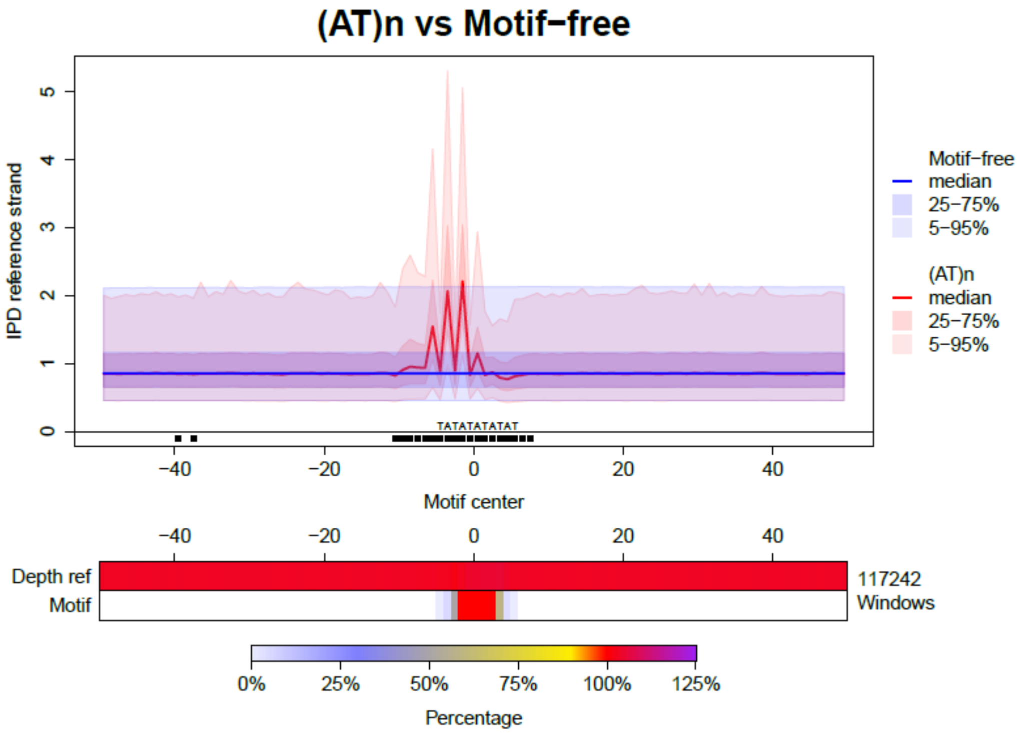
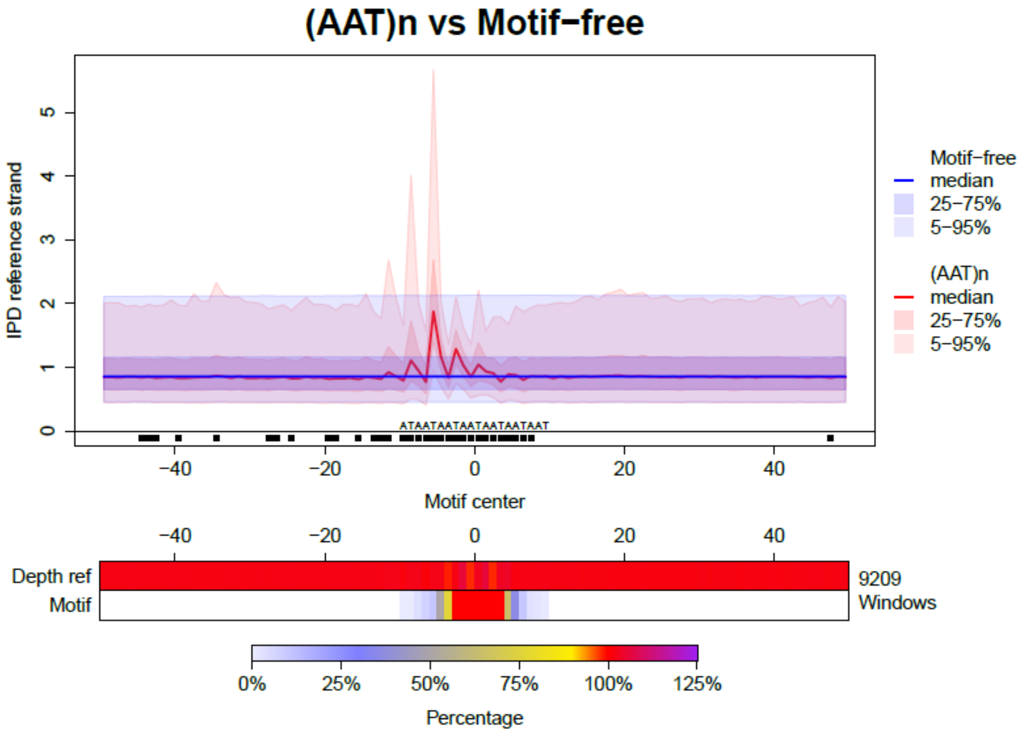

E
Z−DNA vs Motif−free
IPD reference strand
0.0
0.5
1.0
1.5
2.0
2.5
−40
−20
0
20
40
Motif center
Motif−free
median
25−75%
5−95%
Z−DNA
median
25−75%
5−95%
Depth ref
Motif
6207
Windows
0%
25%
50%
75%
100%
125%
Percentage
Z−DNA vs Motif−free
IPD reverse strand
0.0
0.5
1.0
1.5
2.0
2.5
3.0
−40
−20
0
20
40
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.

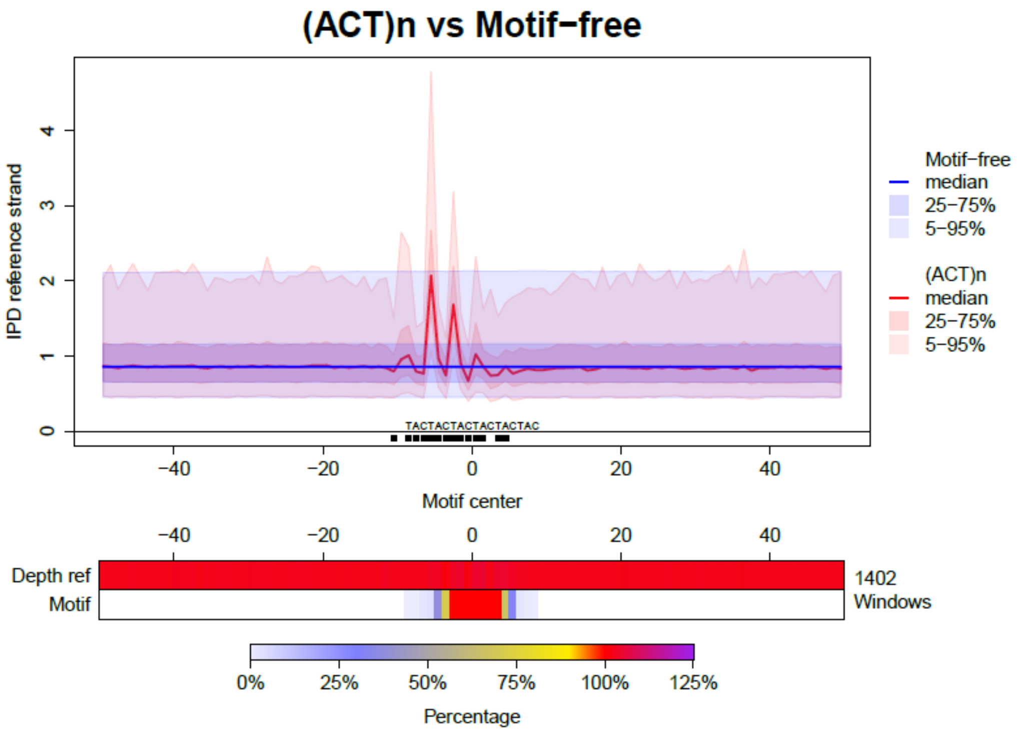
A



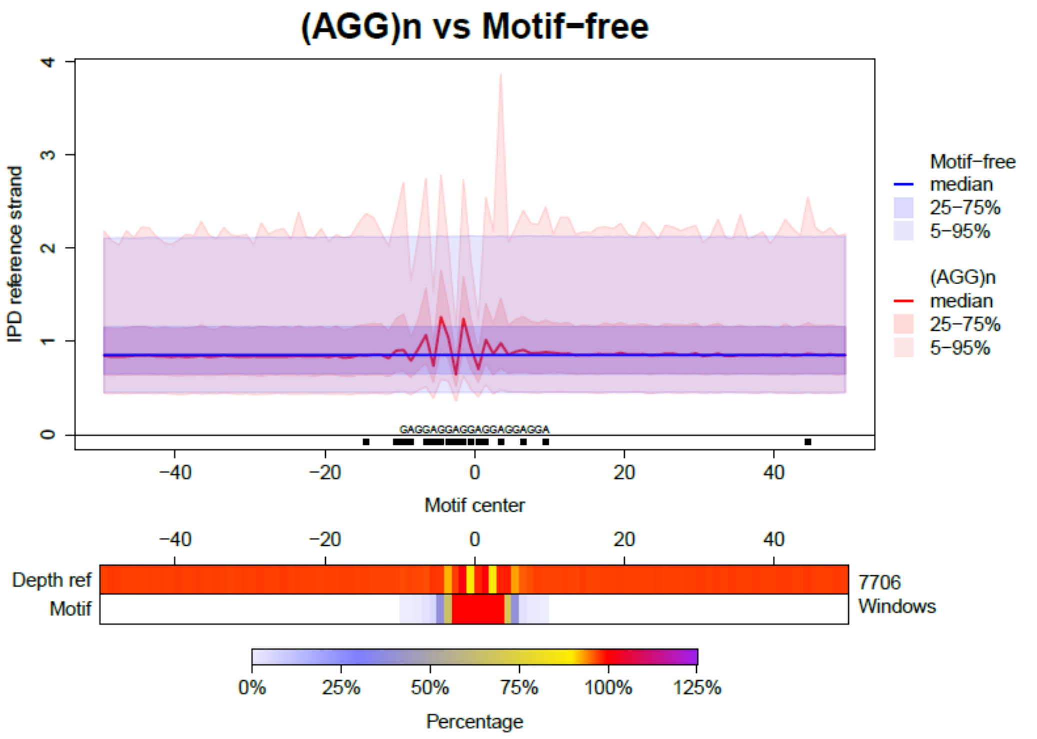
B



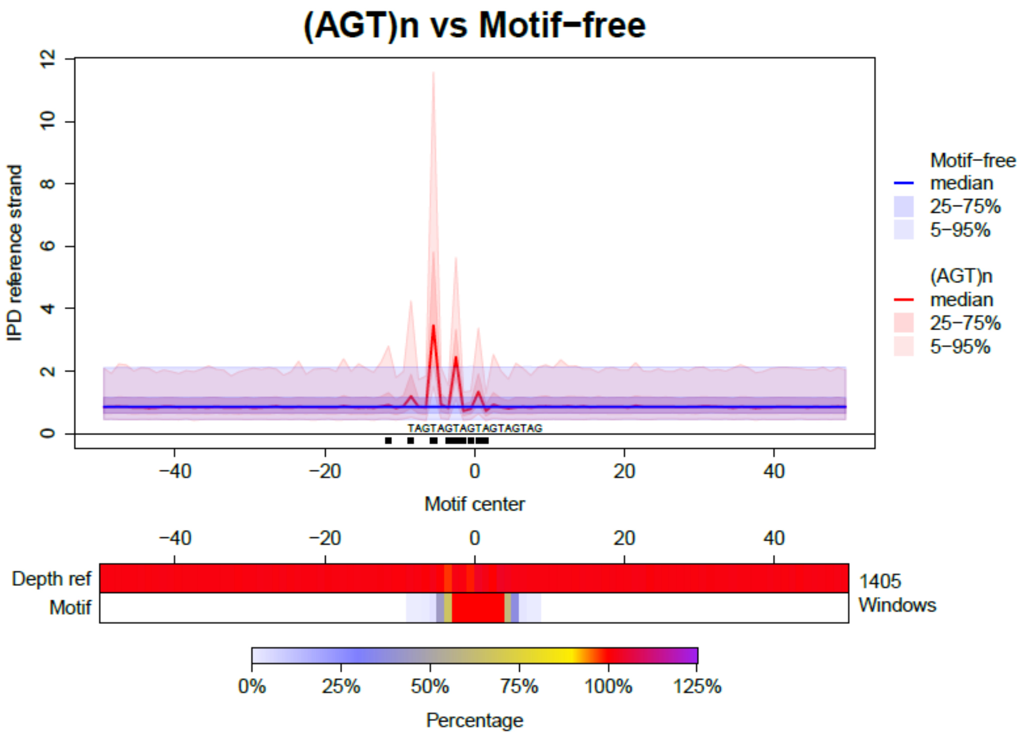
C


(ACT)n vs Motif−free
IPD reference strand
0
1
2
3
4
TACTACTACTACTACTAC
−40
−20
0
20
40
Motif center
Depth ref
Motif
1402
Windows
0%
25%
50%
75%
100%
125%
Percentage
Motif−free
median
25−75%
5−95%
(ACT)n
median
25−75%
5−95%

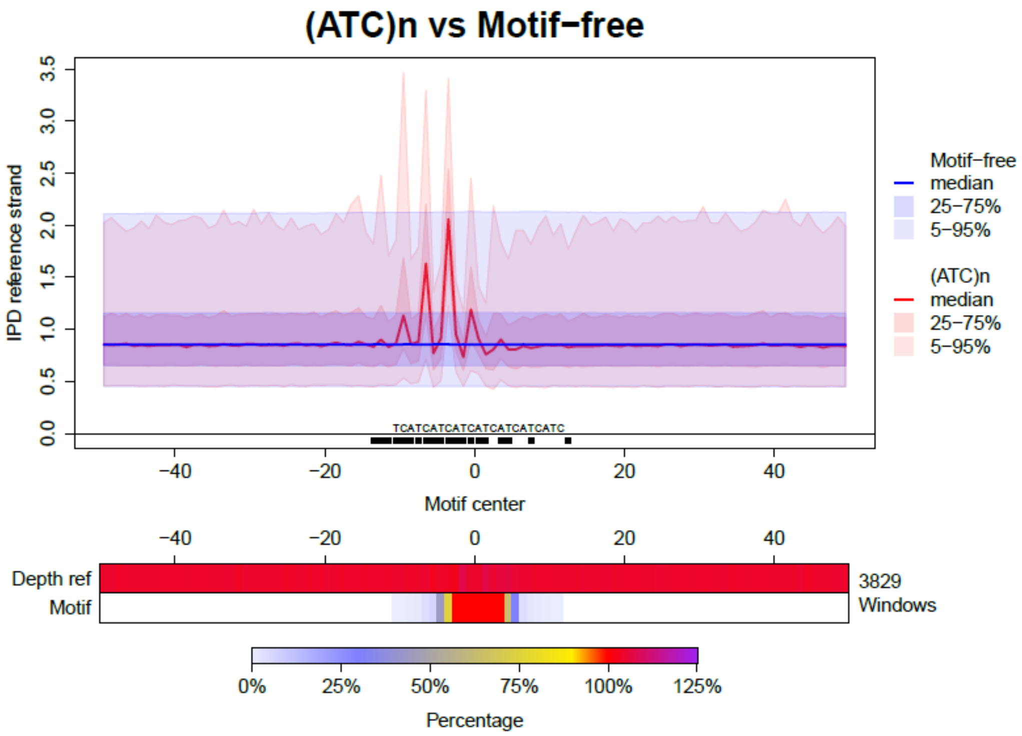
D


(AGG)n vs Motif−free
IPD reference strand
0
1
2
3
4
GAGGAGGAGGAGGAGGAGGA
−40
−20
0
20
40
Motif center
Depth ref
Motif
7706
Windows
0%
25%
50%
75%
100%
125%
Percentage
Motif−free
median
25−75%
5−95%
(AGG)n
median
25−75%
5−95%

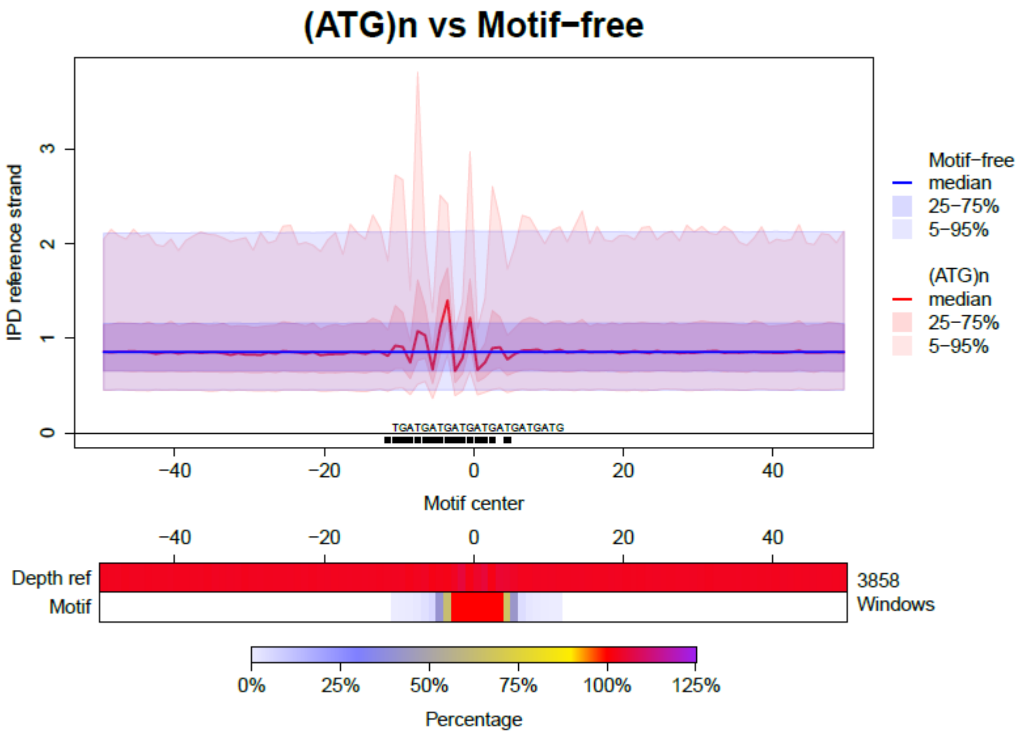
E


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

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
TCATCATCATCATCATCATCATC
−40
−20
0
20
40
Motif center
Motif−free
median
25−75%
5−95%
(ATC)n
median
25−75%
5−95%
Depth ref
Motif
3829
Windows
0%
25%
50%
75%
100%
125%
Percentage

G


(ATG)n vs Motif−free
IPD reference strand
0
1
2
3
TGATGATGATGATGATGATGATG
−40
−20
0
20
40
Motif center
Motif−free
median
25−75%
5−95%
(ATG)n
median
25−75%
5−95%
−40
−20
0
20
40
Depth ref
Motif
3858
Windows
0%
25%
50%
75%
100%
125%
Percentage

H

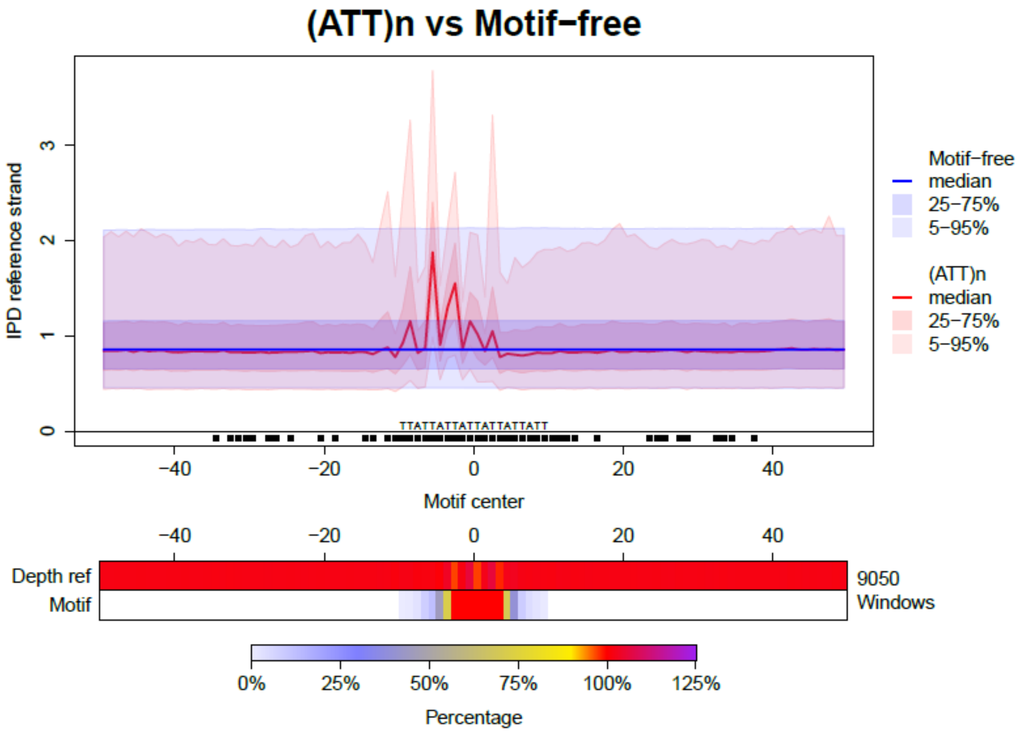
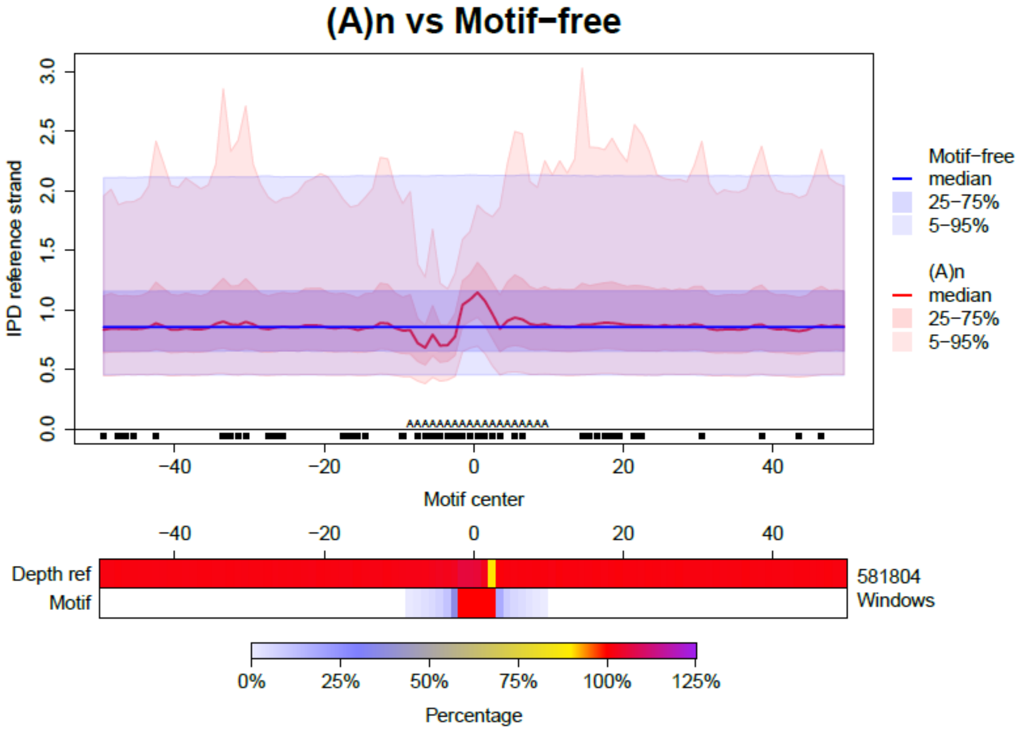

(ATT)n vs Motif−free
IPD reference strand
0
1
2
3
TTATTATTATTATTATTATT
−40
−20
0
20
40
Motif center
Motif−free
median
25−75%
5−95%
(ATT)n
median
25−75%
5−95%
−40
−20
0
20
40
Depth ref
Motif
9050
Windows
0%
25%
50%
75%
100%
125%
Percentage

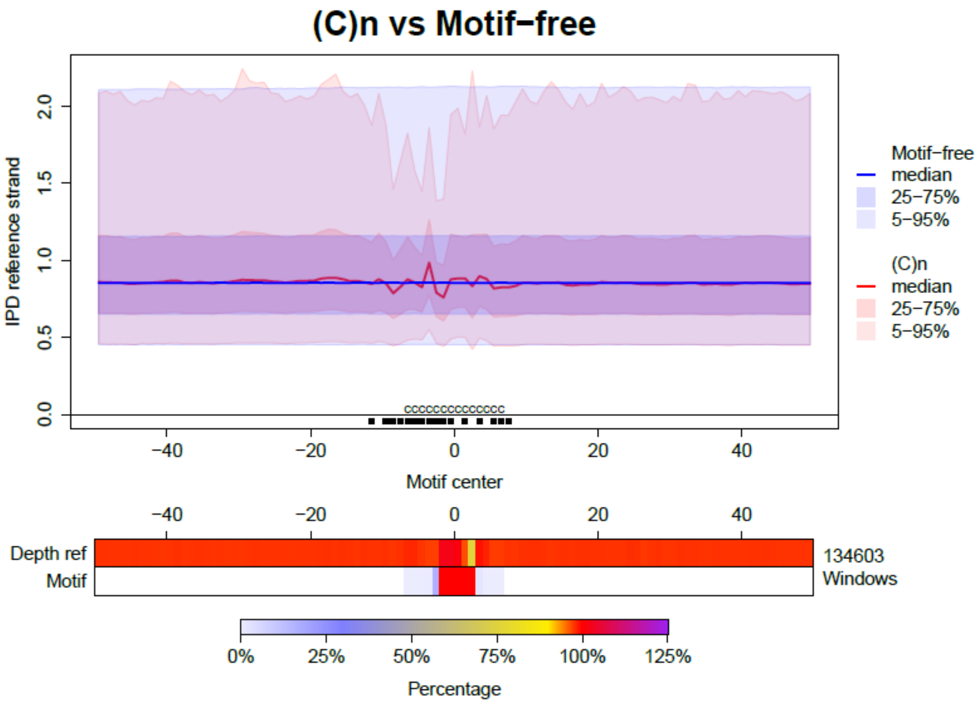
Figure S8. The effect of homopolymers and STRs that can form H-DNA on polymerization kinetics.

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).

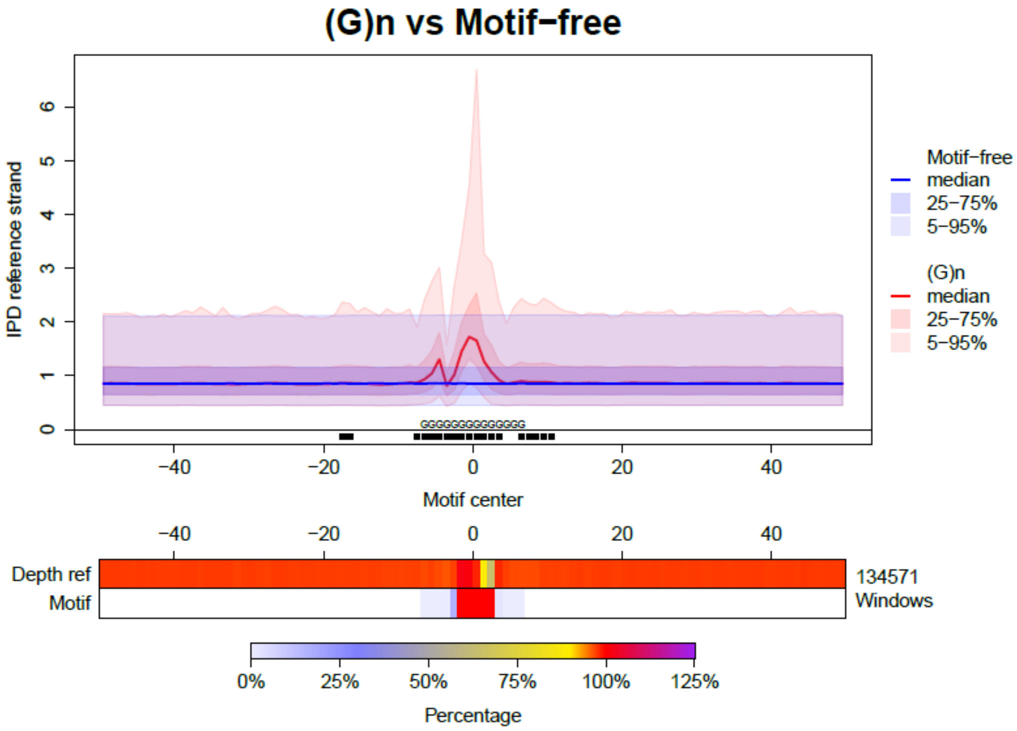
A



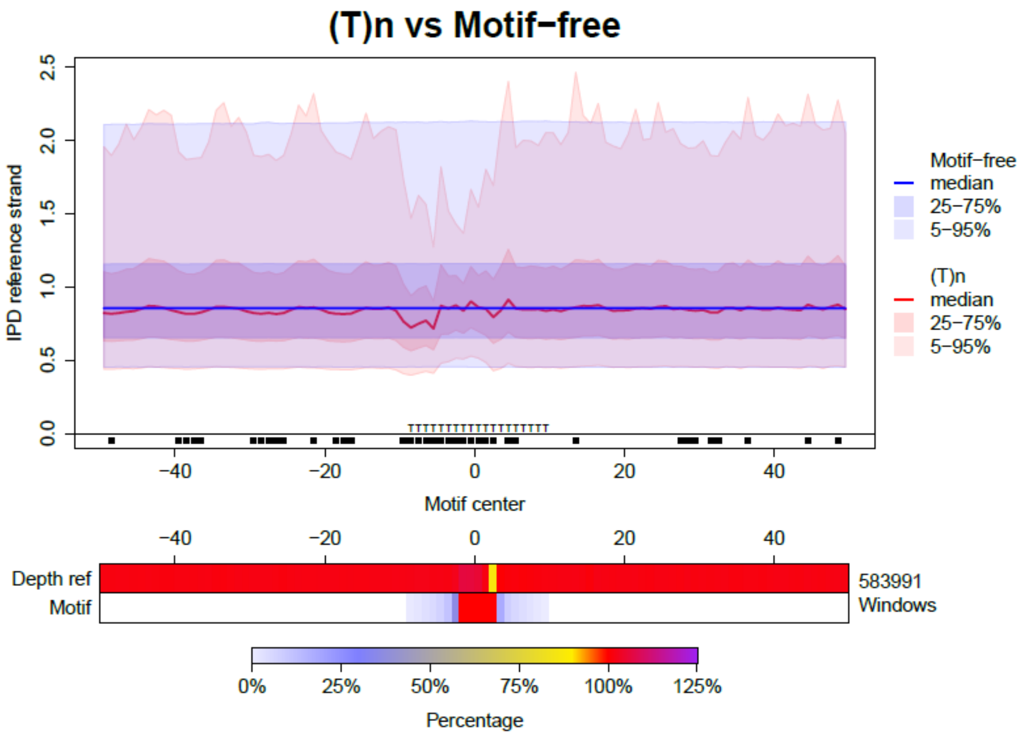
B


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

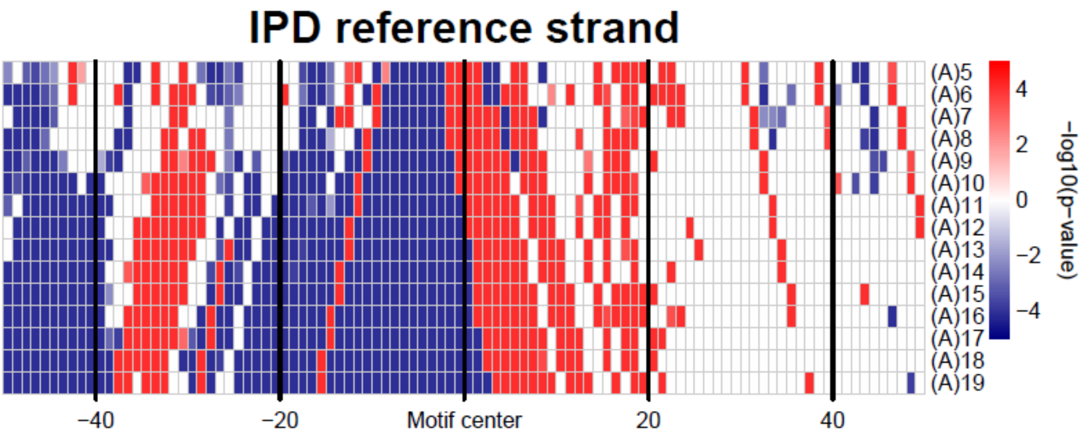
C


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

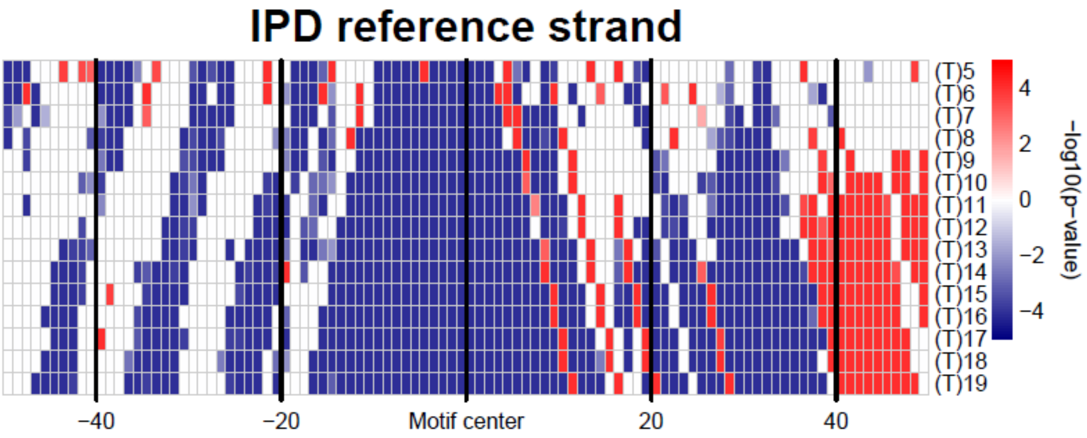
D


(T)n vs Motif−free
IPD reference strand
2.5
2.0
1.5
1.0
0.5
0.0
−40
−20
0
20
40
Motif center
Motif−free
median
25−75%
5−95%
(T)n
median
25−75%
5−95%
TTTTTTTTTTTTTTTTTTT
Depth ref
Motif
583991
Windows
0%
25%
50%
75%
100%
125%
Percentage

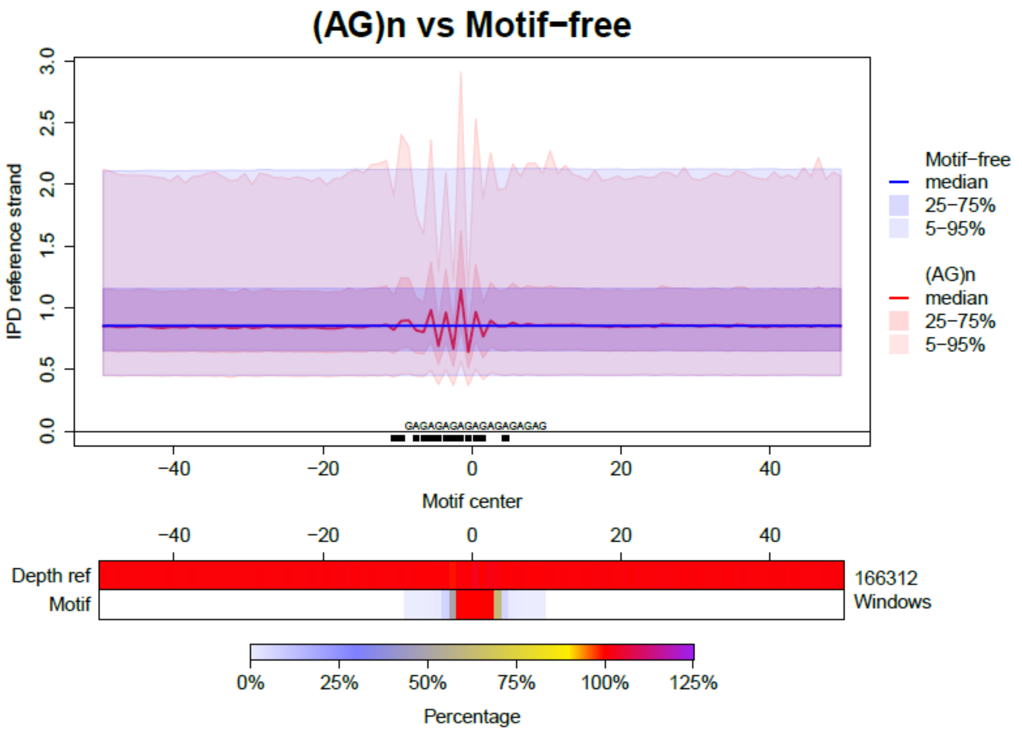
E


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
4
2
0
−2
−4
−log10(p−value)
−40
−20
Motif center
20
40

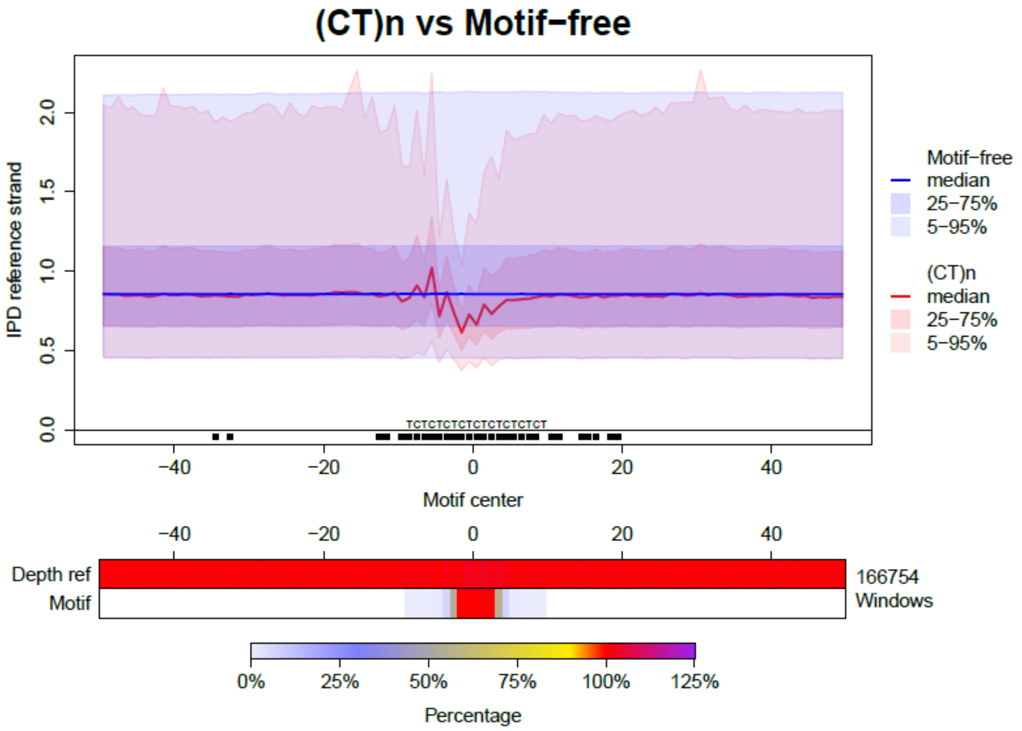
F


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
4
2
0
−2
−4
−log10(p−value)
−40
−20
Motif center
20
40

G


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

H


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

I

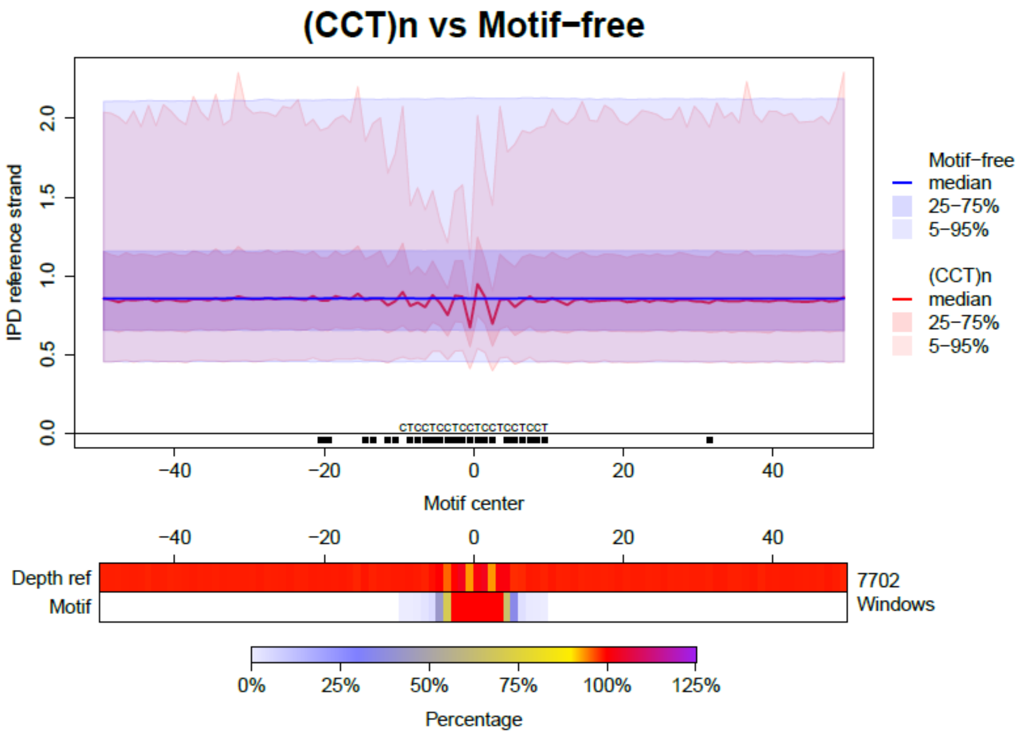
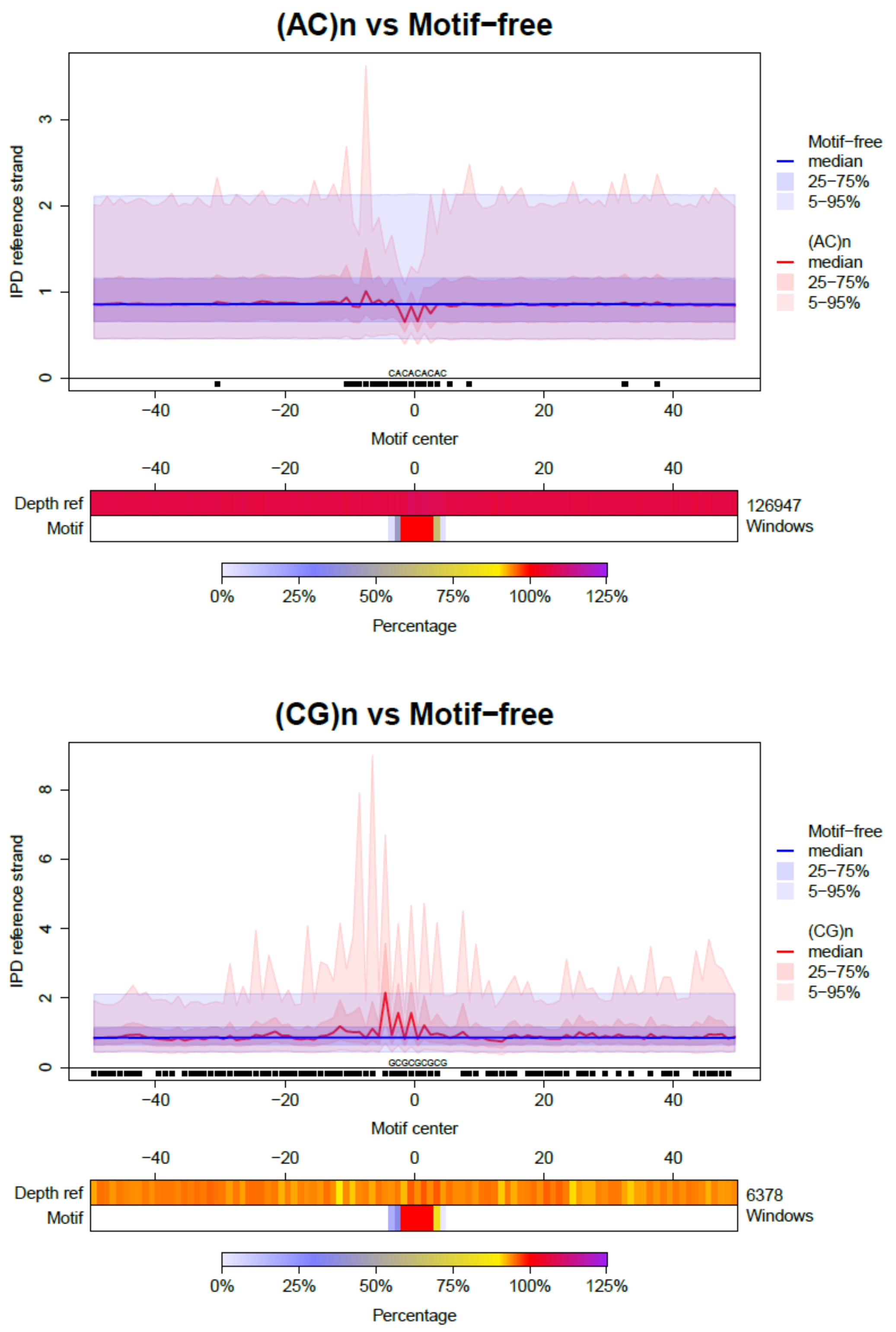

(CCT)n vs Motif−free
IPD reference strand
0.0
0.5
1.0
1.5
2.0
CTCCTCCTCCTCCTCCTCCT
−40
−20
0
20
40
Motif center
Motif−free
median
25−75%
5−95%
(CCT)n
median
25−75%
5−95%
−40
−20
0
20
40
Depth ref
Motif
7702
Windows
0%
25%
50%
75%
100%
125%
Percentage

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



C

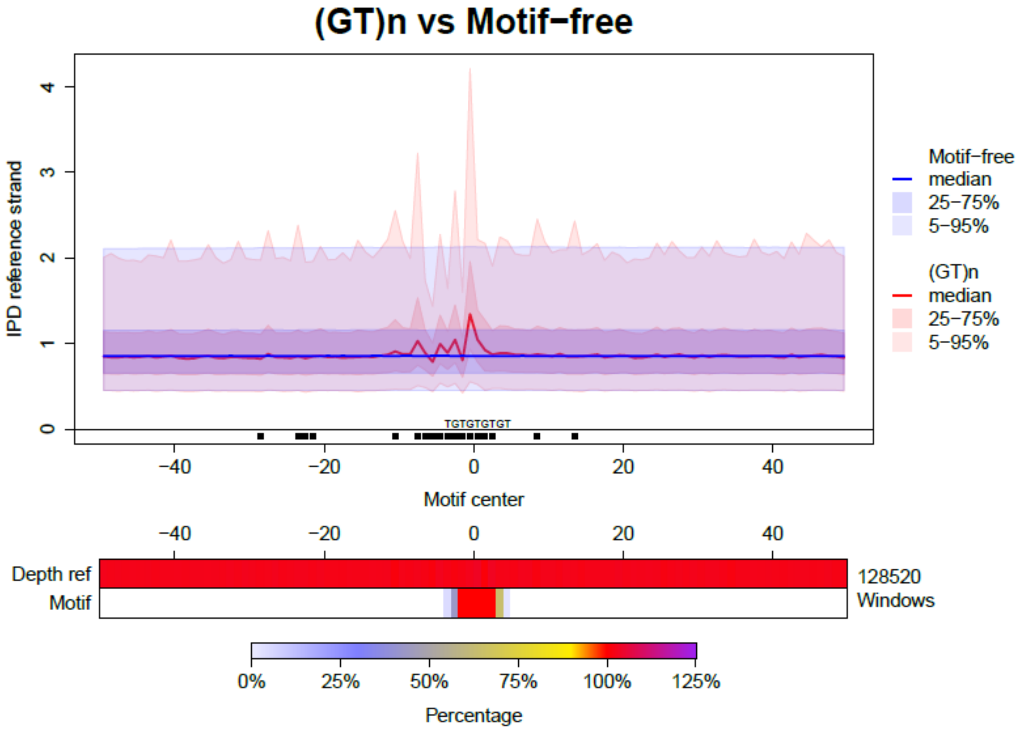
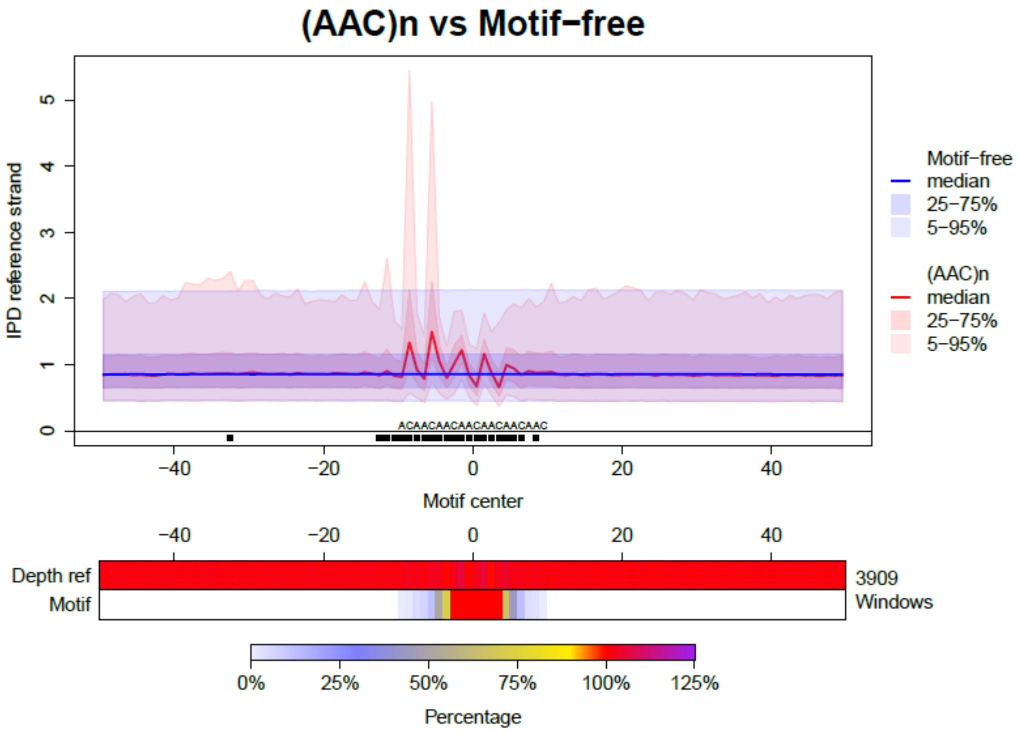

(GT)n vs Motif−free
IPD reference strand
0
1
2
3
4
TGTGTGTGT
−40
−20
0
20
40
Motif center
Motif−free
median
25−75%
5−95%
(GT)n
median
25−75%
5−95%
Depth ref
Motif
128520
Windows
0%
25%
50%
75%
100%
125%
Percentage

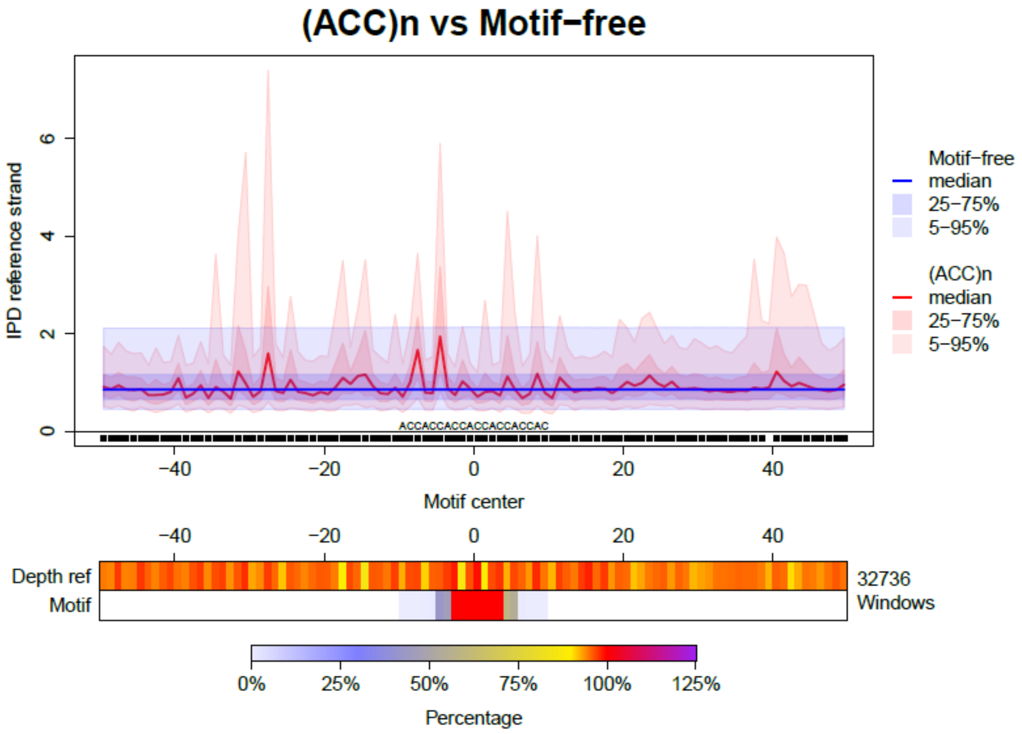
Figure S10. The effect of STRs on polymerization kinetics.

A (AAC)$_n$. **B** (ACC)$_n$. **C** (ACG)$_n$. **D** (CGT)$_n$. **E** (GGT)$_n$. **F** (GTT)$_n$. See the legend of Fig. 2A.

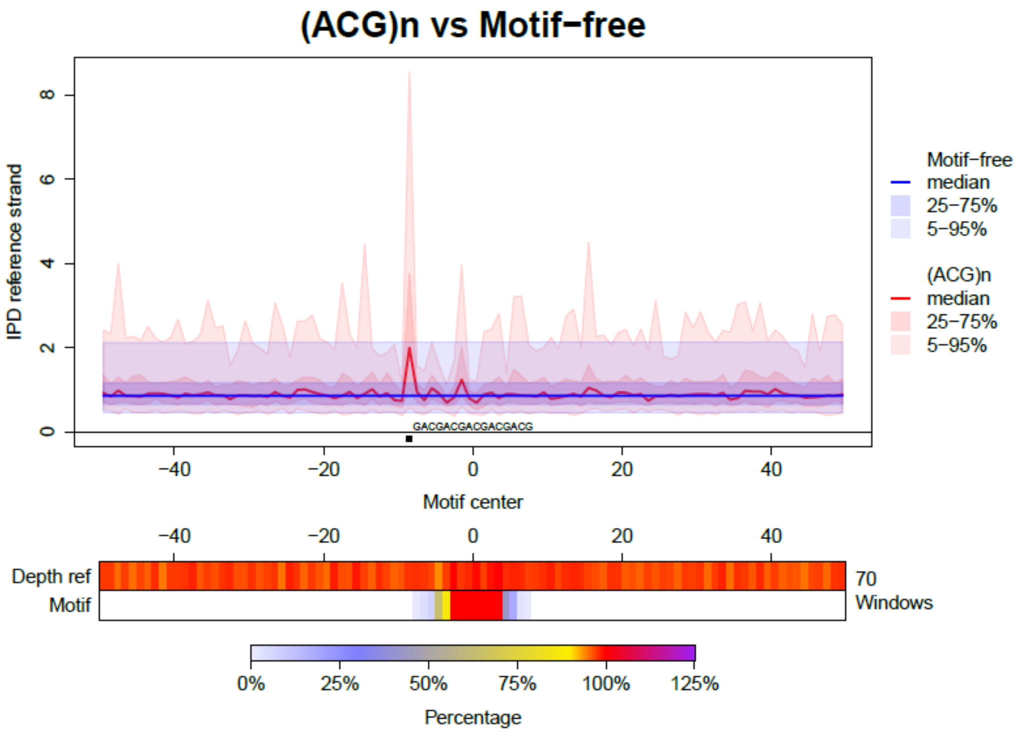
A



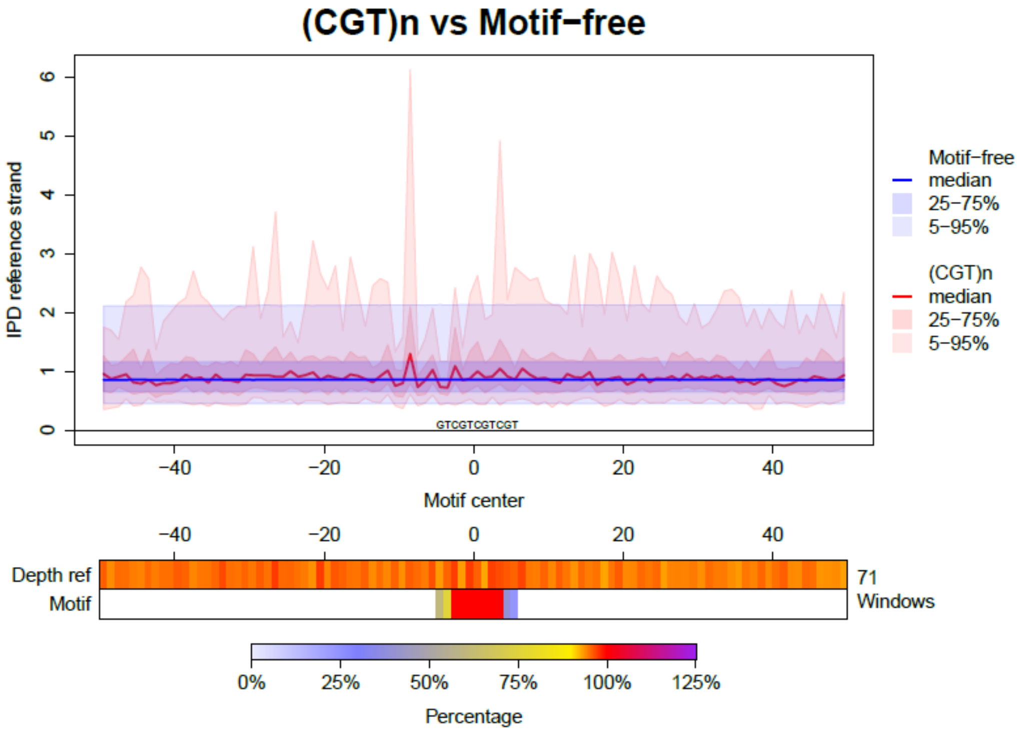
B



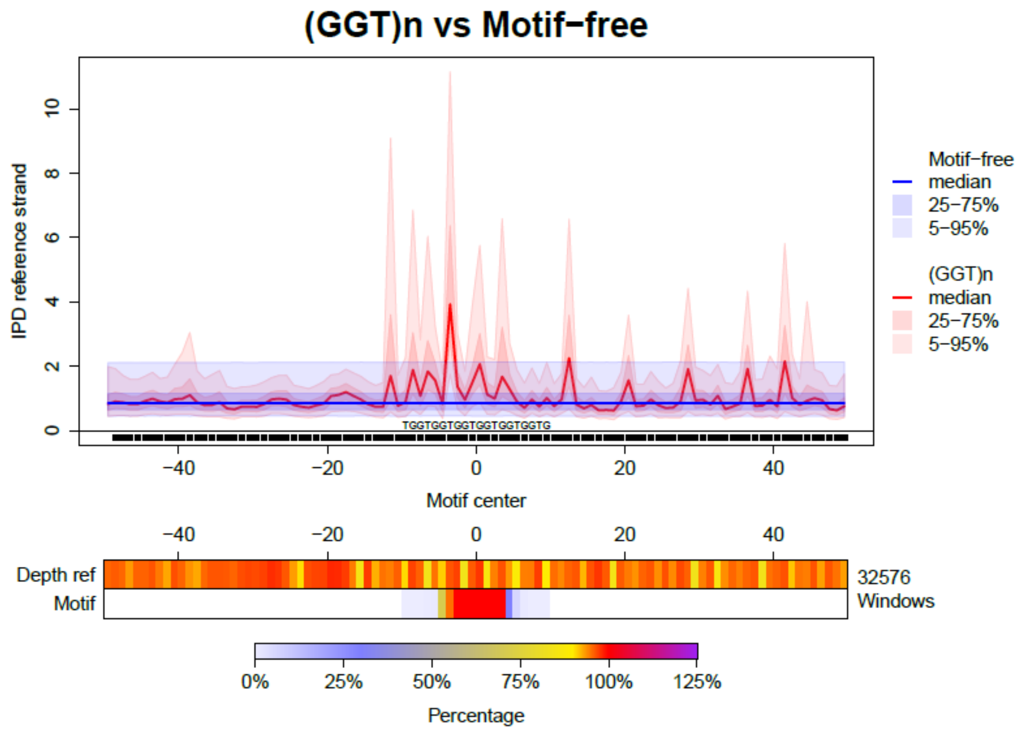
C


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

D


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

E


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

F

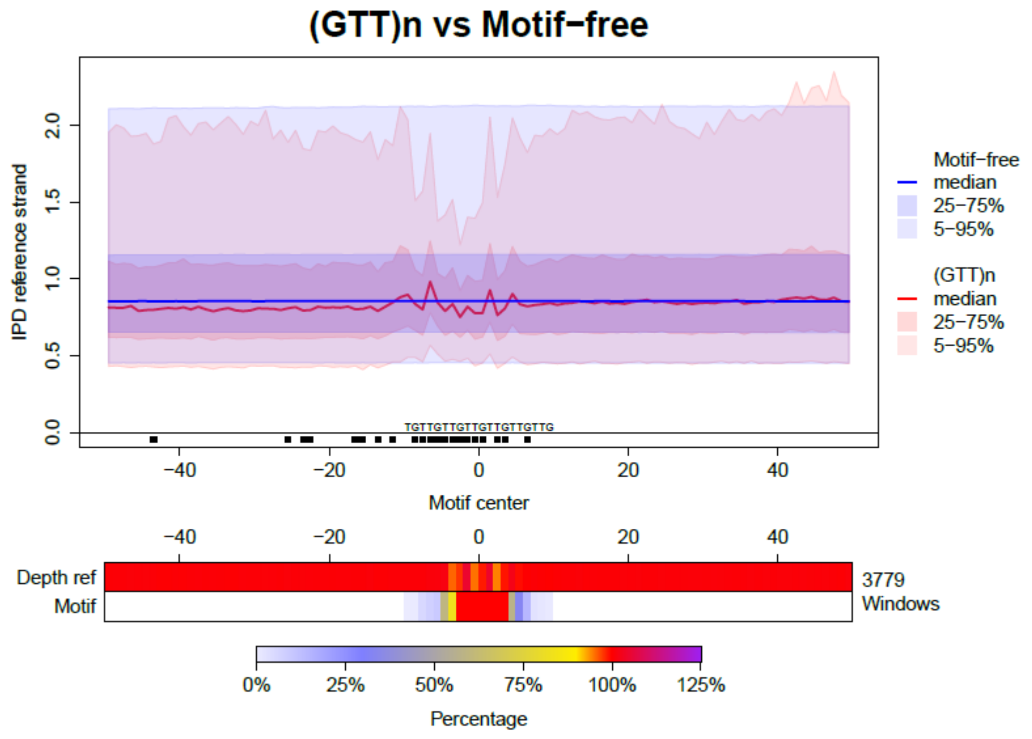
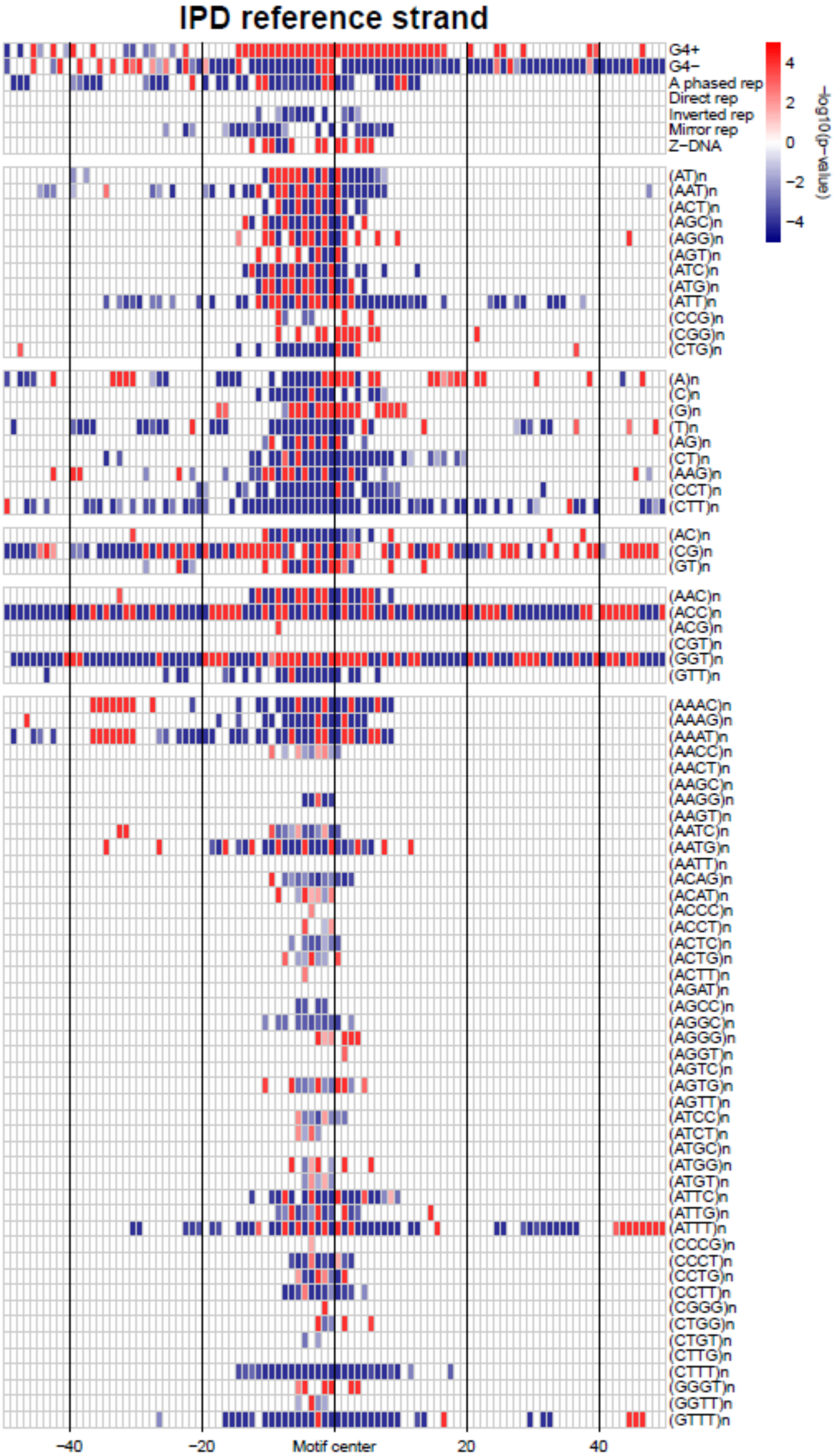
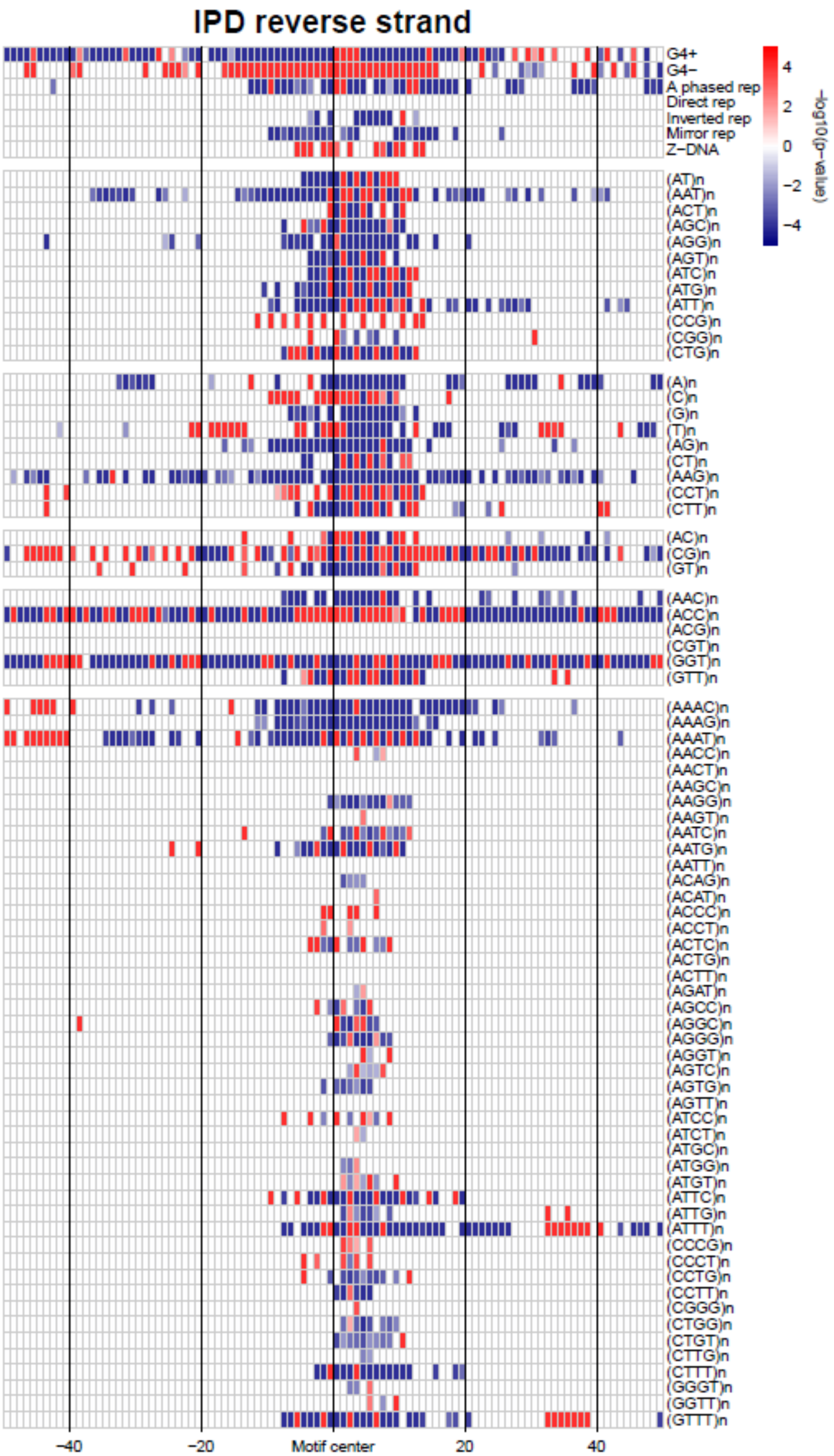

(GTT)n vs Motif−free
IPD reference strand
0.0
0.5
1.0
1.5
2.0
TGTTGTTGTTGTTGTTGTTG
−40
−20
0
20
40
Motif center
Motif−free
median
25−75%
5−95%
(GTT)n
median
25−75%
5−95%
−40
−20
0
20
40
Depth ref
Motif
3779
Windows
0%
25%
50%
75%
100%
125%
Percentage

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.

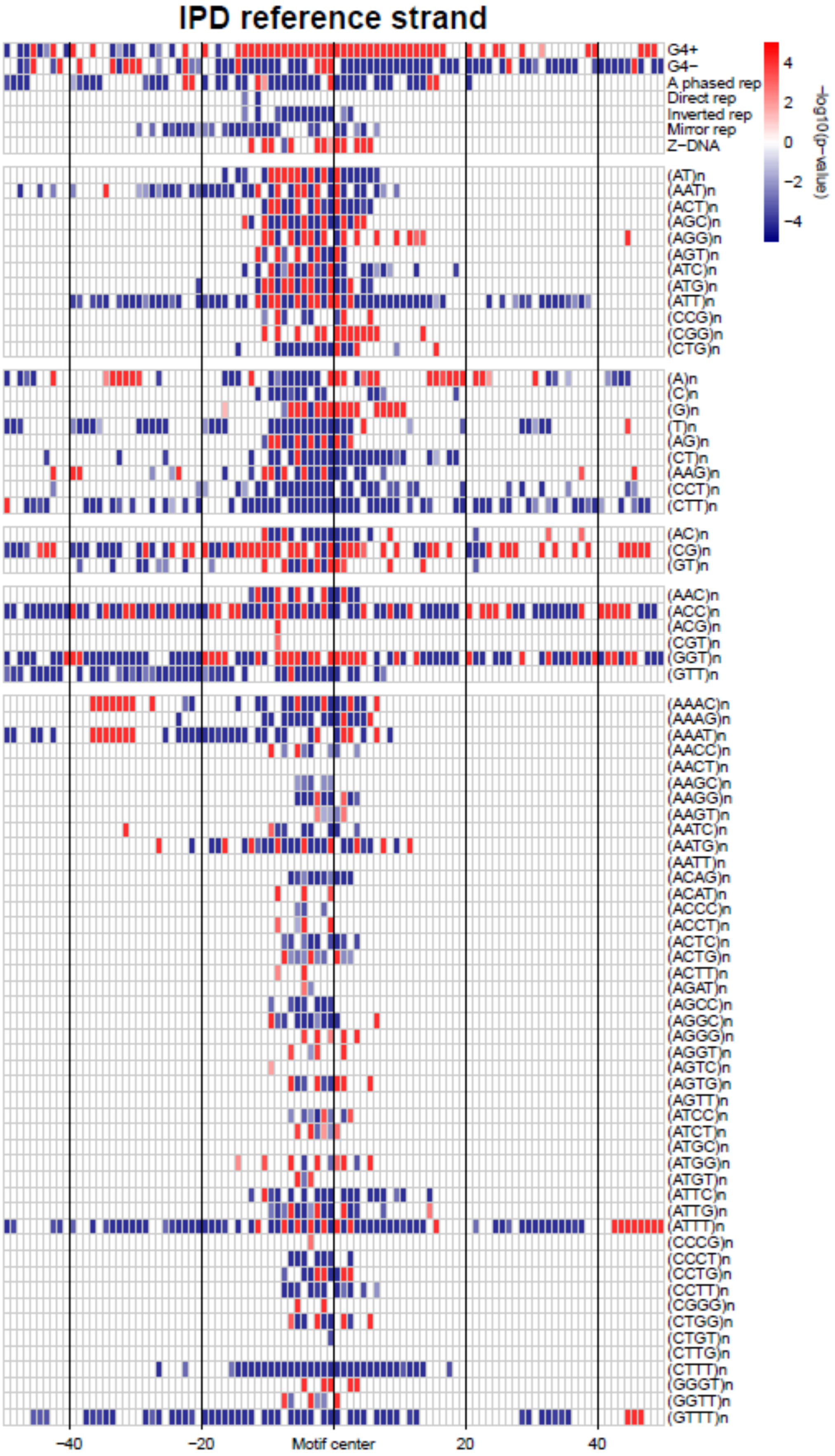
A


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
−40
−20
Motif center
20
40
4
2
0
−2
−4
−log10(p−value)

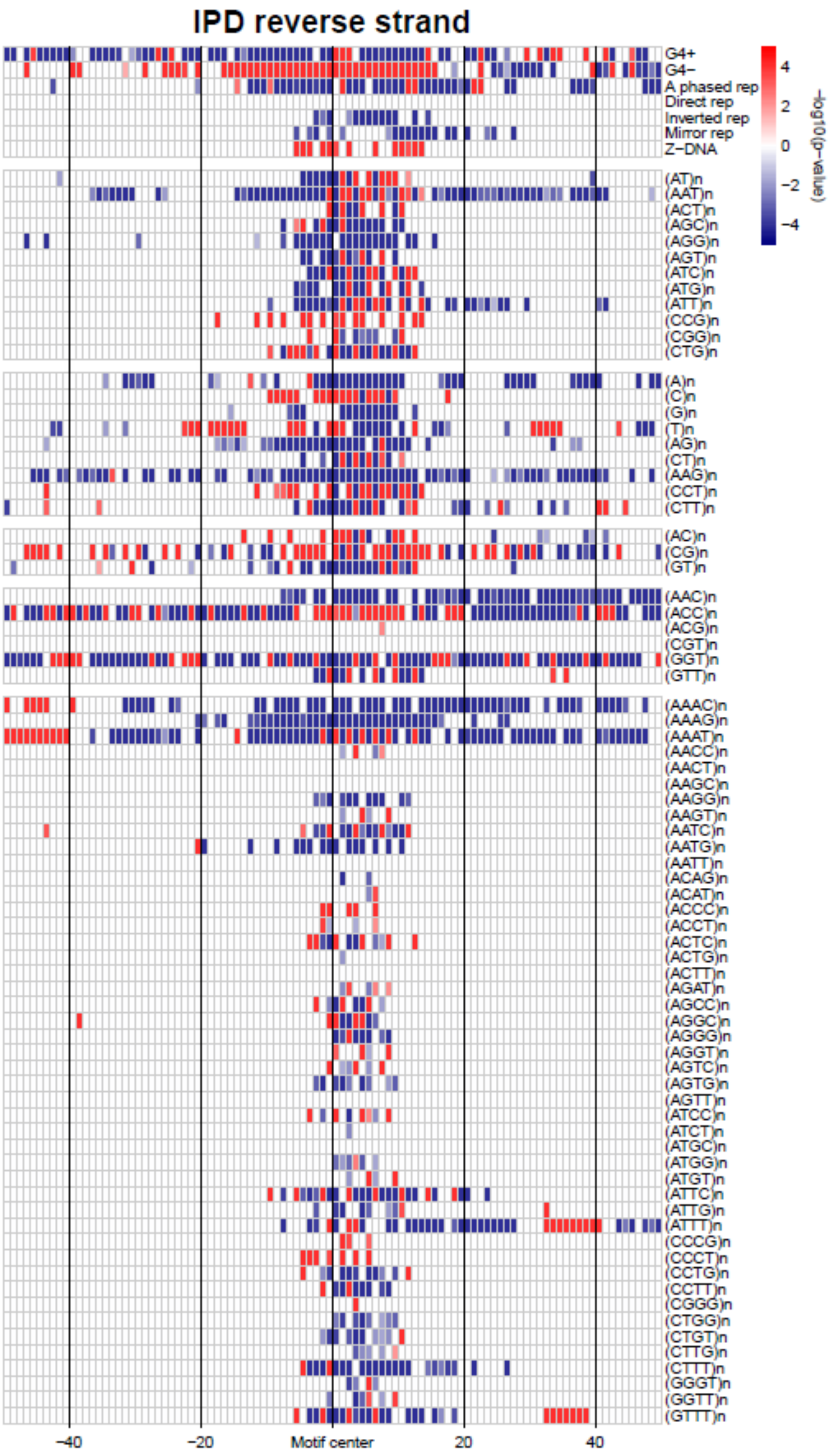
B


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
−40
−20
Motif center
20
40
4
2
0
−2
−4
−log10(p−value)

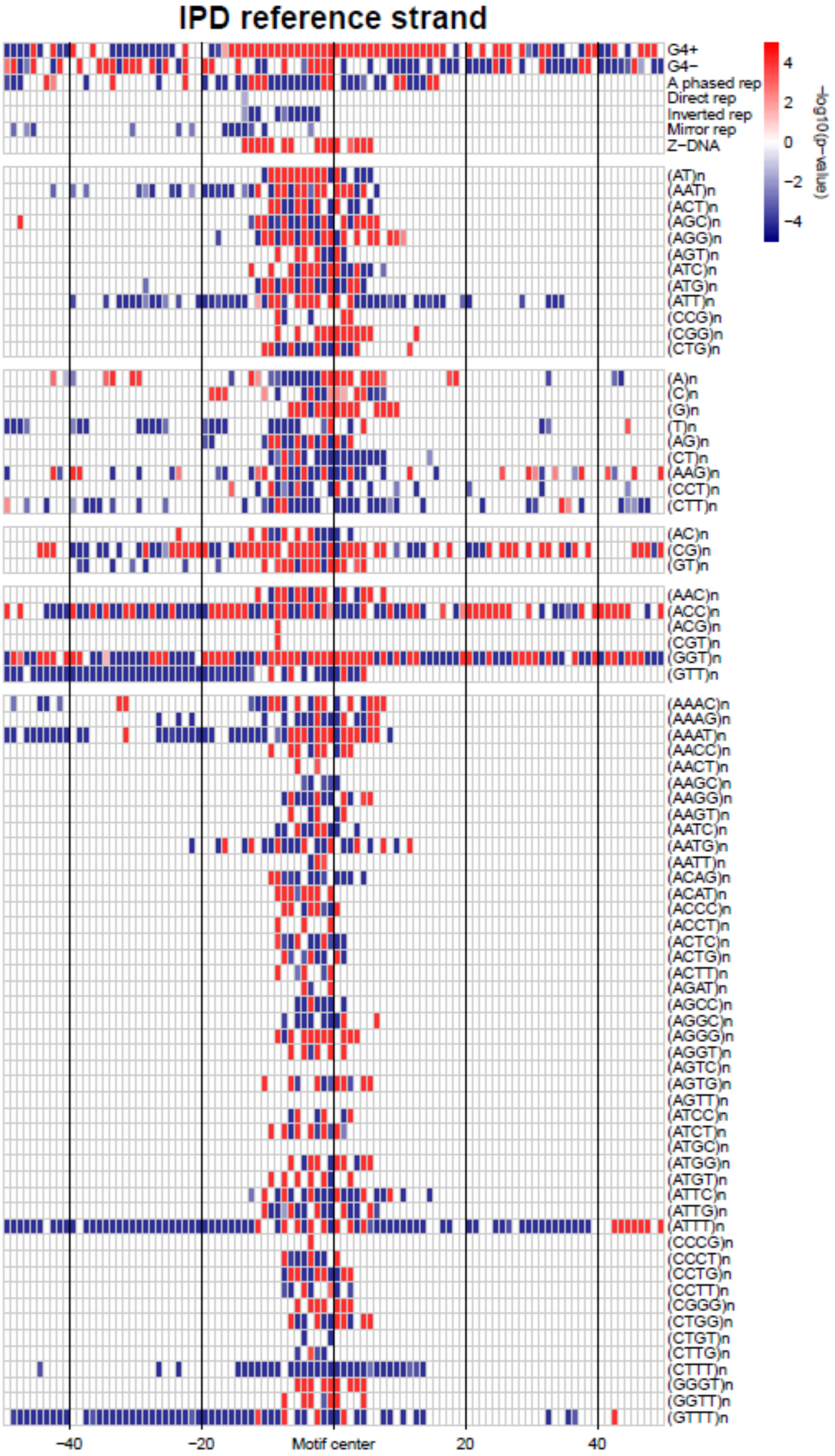
C


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
−40
−20
Motif center
20
40
4
2
0
−2
−4
−log10(p−value)

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
−40
−20
Motif center
20
40
4
2
0
−2
−4
−log10(p−value)

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
−40
−20
Motif center
20
40
4
2
0
−2
−4
−log10(p−value)

F

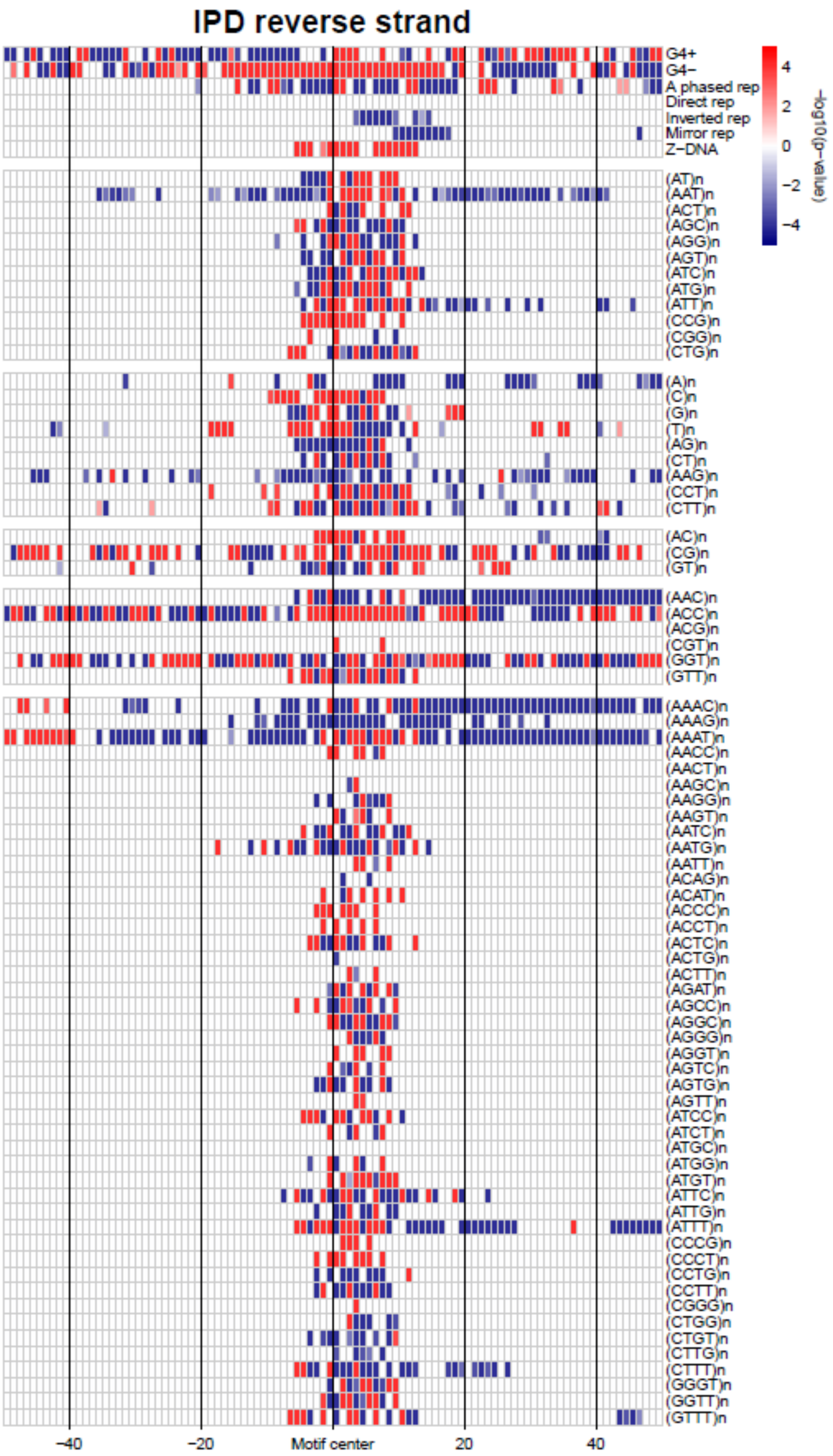
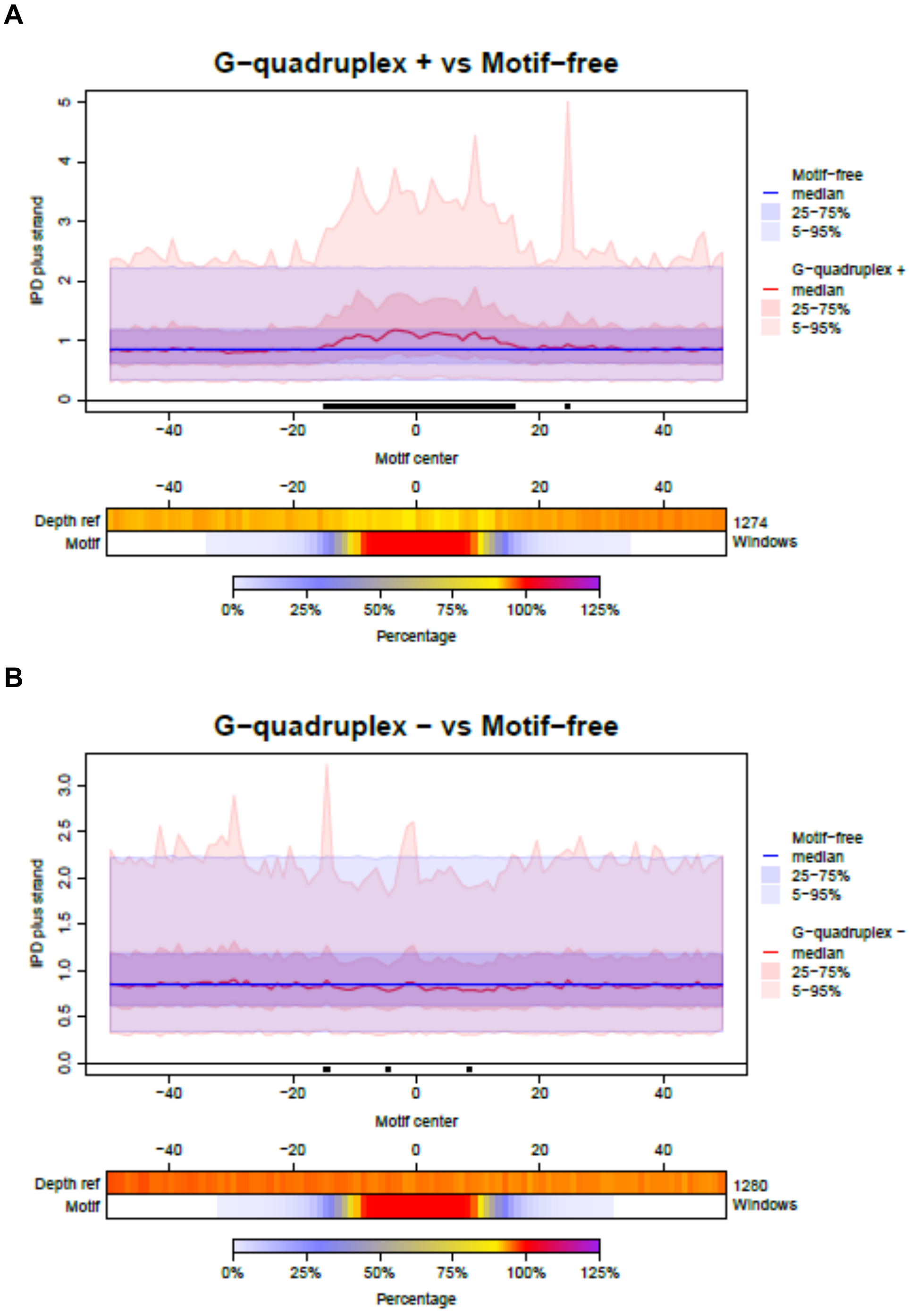
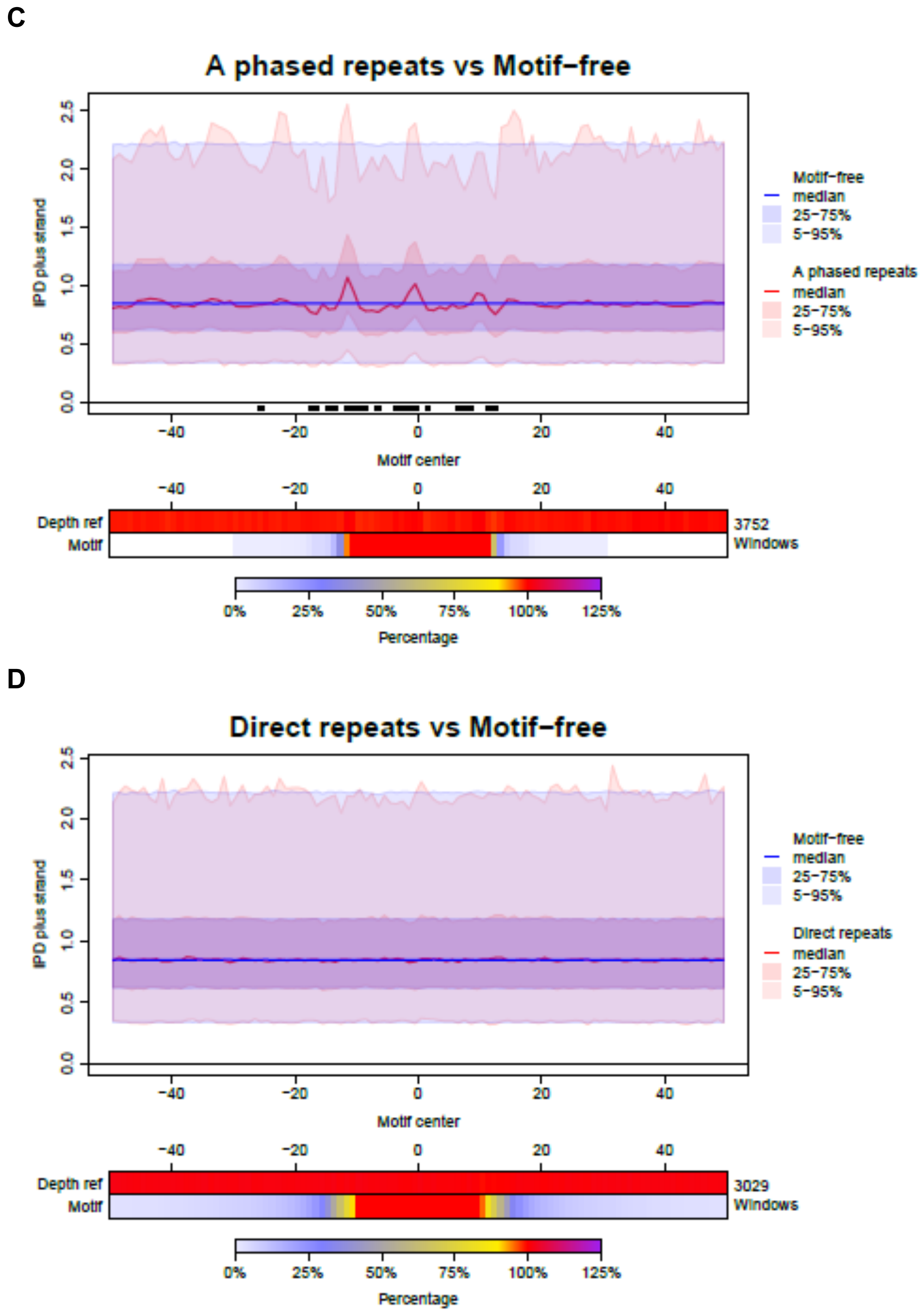
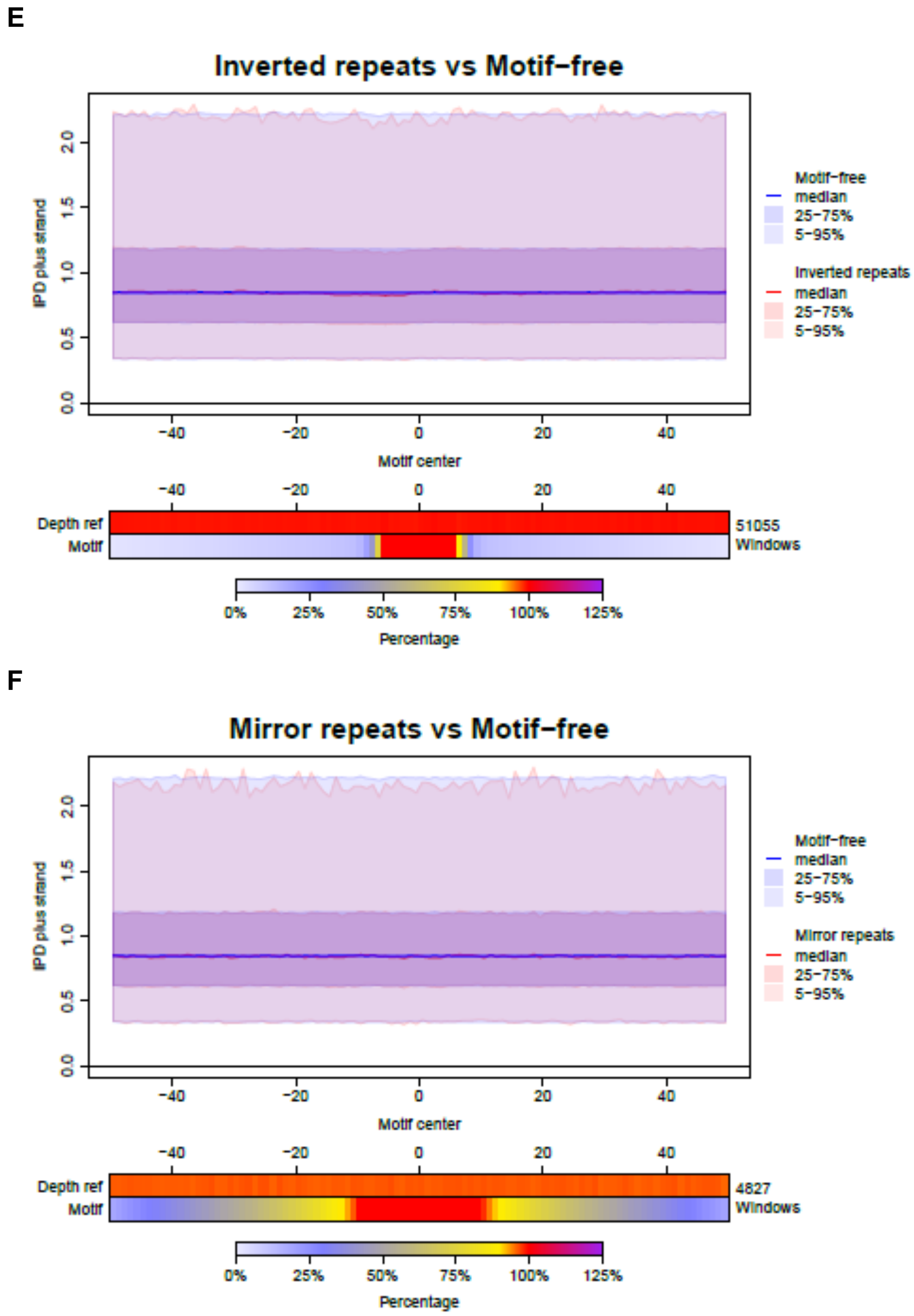

IPD reverse strand
G4+
G4−
A phased rep
Direct rep
Inverted rep
Mirror rep
Z−DNA
(AT)n
(AAT)n
(ACT)n
(AGC)n
(AGG)n
(AGT)n
(ATC)n
(ATG)n
(ATT)n
(CCG)n
(CGG)n
(CTG)n
(A)n
(C)n
(G)n
(T)n
(AG)n
(CT)n
(AAG)n
(CCT)n
(CTT)n
(AC)n
(CG)n
(GT)n
(AAC)n
(ACC)n
(ACG)n
(CGT)n
(GGT)n
(GTT)n
(AAAC)n
(AAAG)n
(AAAT)n
(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
−40
−20
Motif center
20
40
4
2
0
−2
−4
−log10(p−value)

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(Yang et al. 1995). Subsequently, the flow-sorted material was used as a template for WGA performed with the REPLI-g Single Cell Kit (Qiagen). After de-branching(Zhang et al. 2006), 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 (Cer et al. 2013). **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.




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

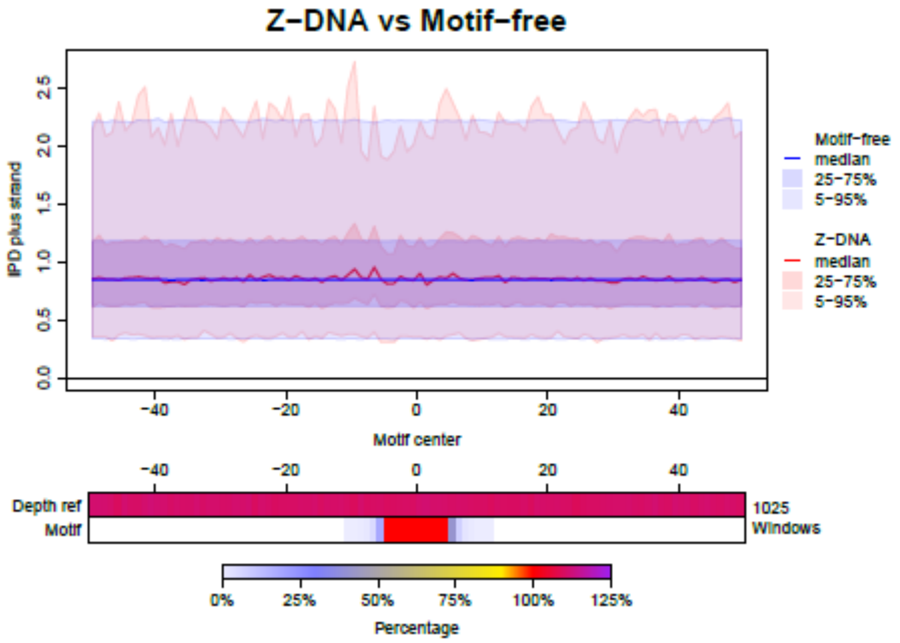
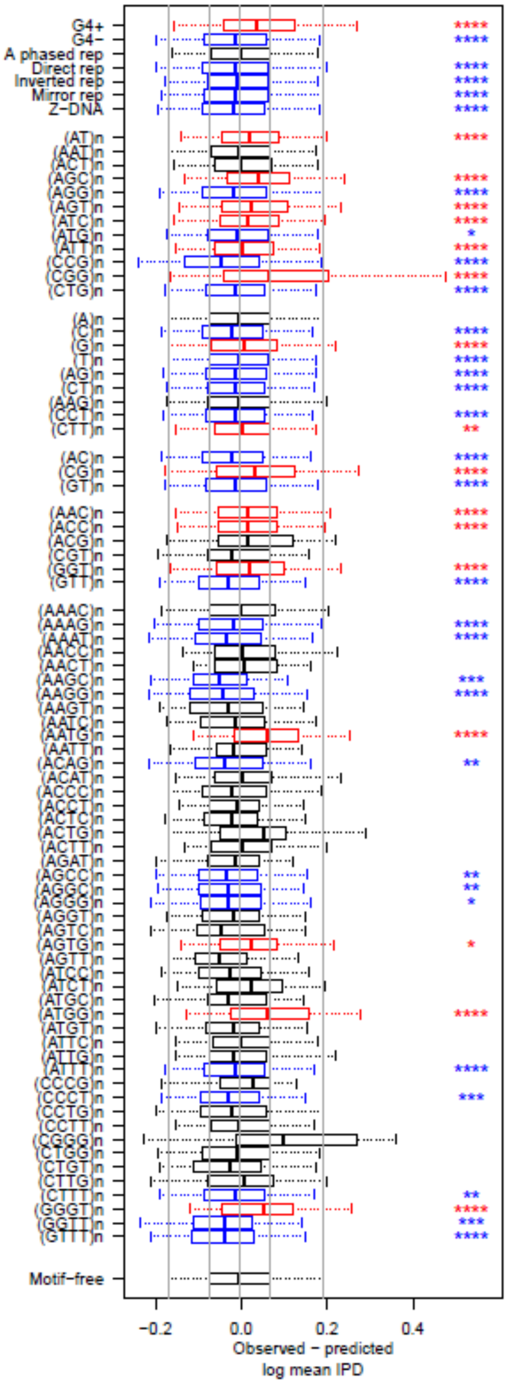

Z−DNA vs Motif−free
IPD plus strand
0.0
0.5
1.0
1.5
2.0
2.5
−40
−20
0
20
40
Motif center
Motif−free
median
25−75%
5−95%
Z−DNA
median
25−75%
5−95%
−40
−20
0
20
40
Depth ref
Motif
1025
Windows
0%
25%
50%
75%
100%
125%
Percentage

Figure S13. 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. Bonferroni-corrected *t*-test p-values for differences: ≤0.0001 '****', ≤0.001 '***', ≤0.01 '**', ≤0.05 '*'. Black: non-significant (corrected p-value > 0.05); red/blue: significant, with observed mean IPDs higher/lower than composition-based predictions. Boxplot whiskers: 5th and 95th quantiles of the differences.

A



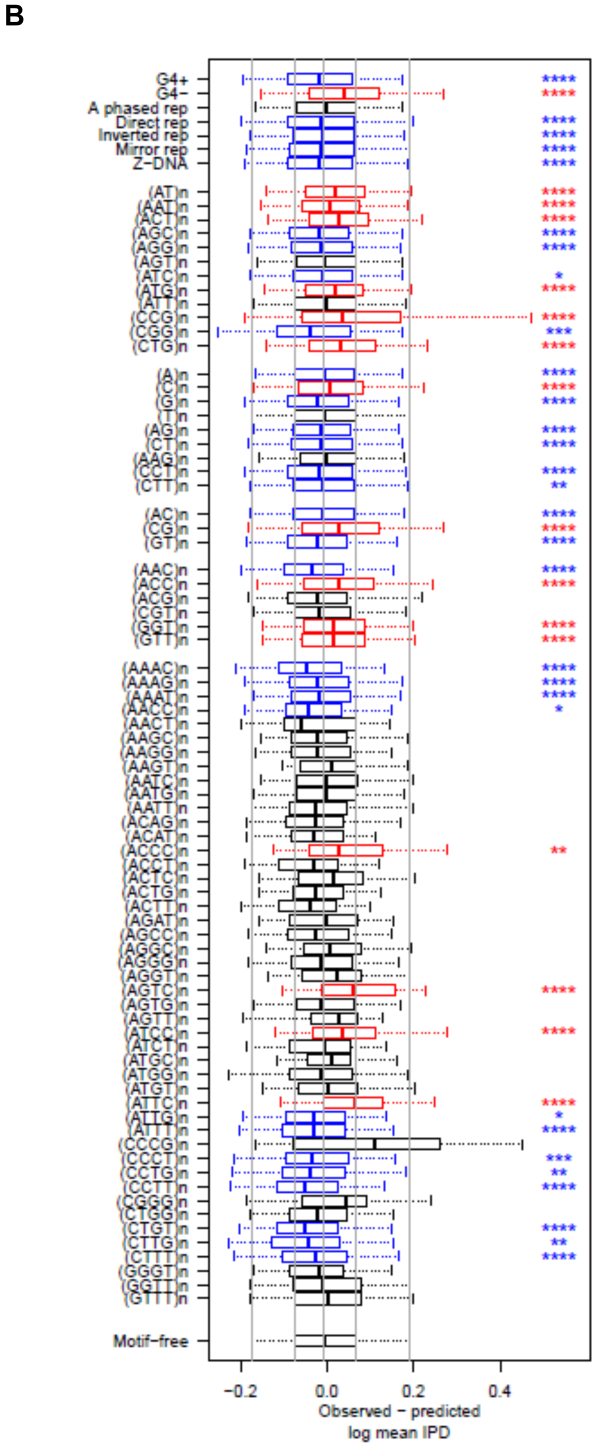

B
G4+
G4−
A phased rep
Direct rep
Inverted rep
Mirror rep
Z−DNA
(AT)n
(AAT)n
(ACT)n
(AGC)n
(AGG)n
(AGT)n
(ATC)n
(ATG)n
(ATT)n
(CCG)n
(CGG)n
(CTG)n
(A)n
(C)n
(G)n
(T)n
(AG)n
(CT)n
(AAG)n
(CCT)n
(CTT)n
(AC)n
(CG)n
(GT)n
(AAC)n
(ACC)n
(ACG)n
(CGT)n
(GGT)n
(GTT)n
(AAAC)n
(AAAG)n
(AAAT)n
(AACC)n
(AACT)n
(AAGC)n
(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 S14. 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.

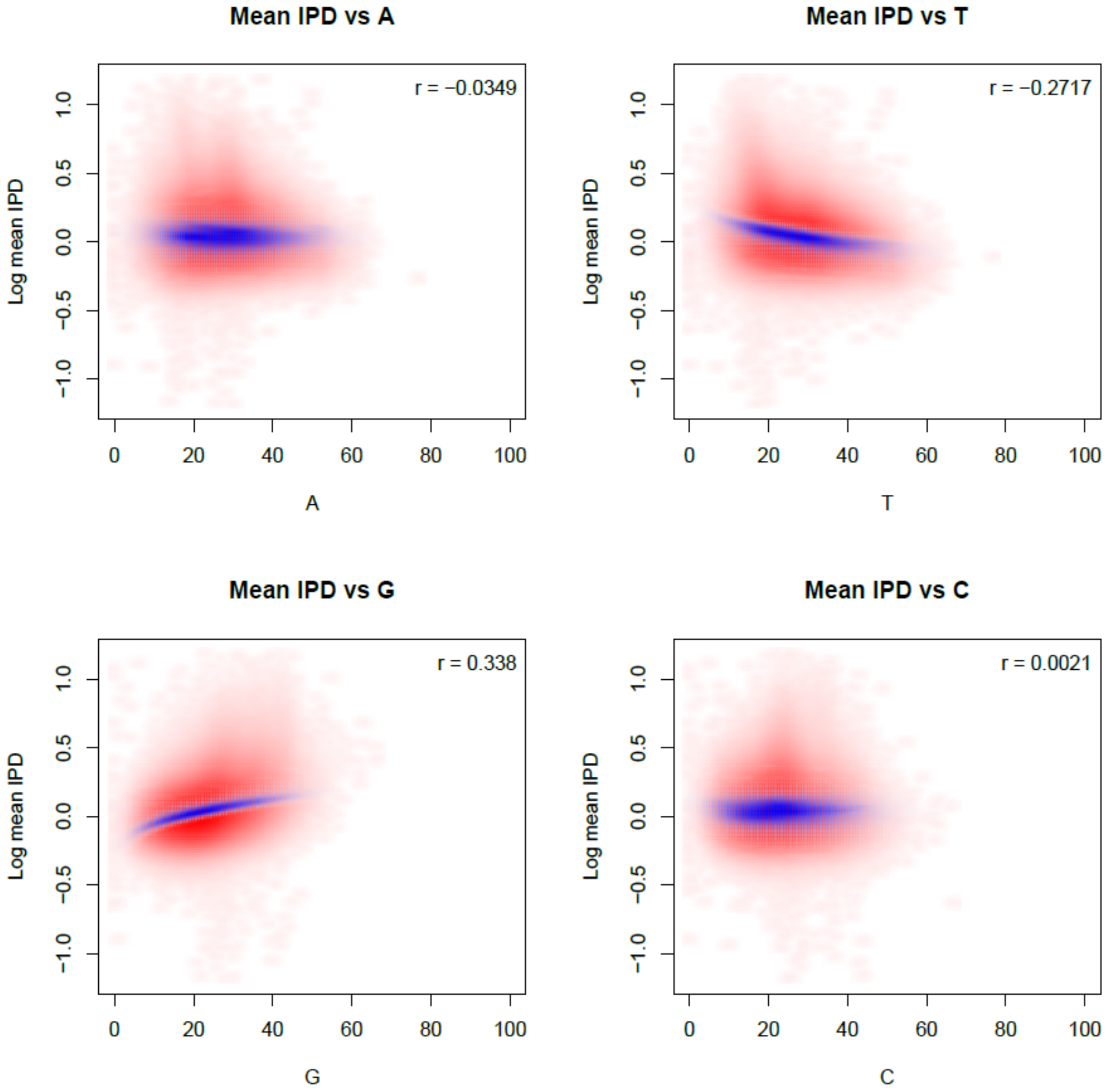
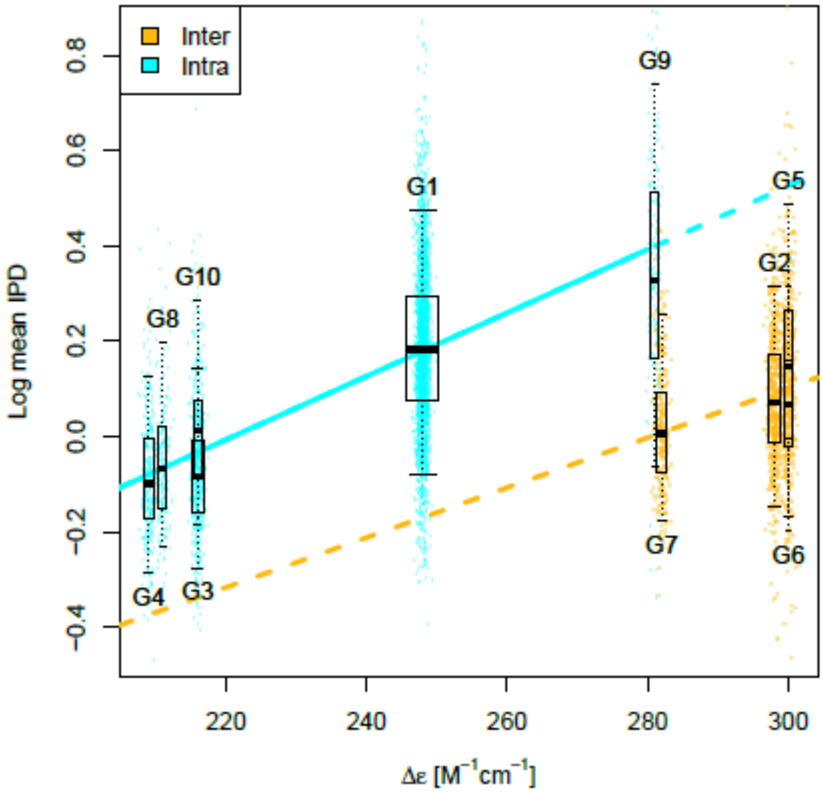


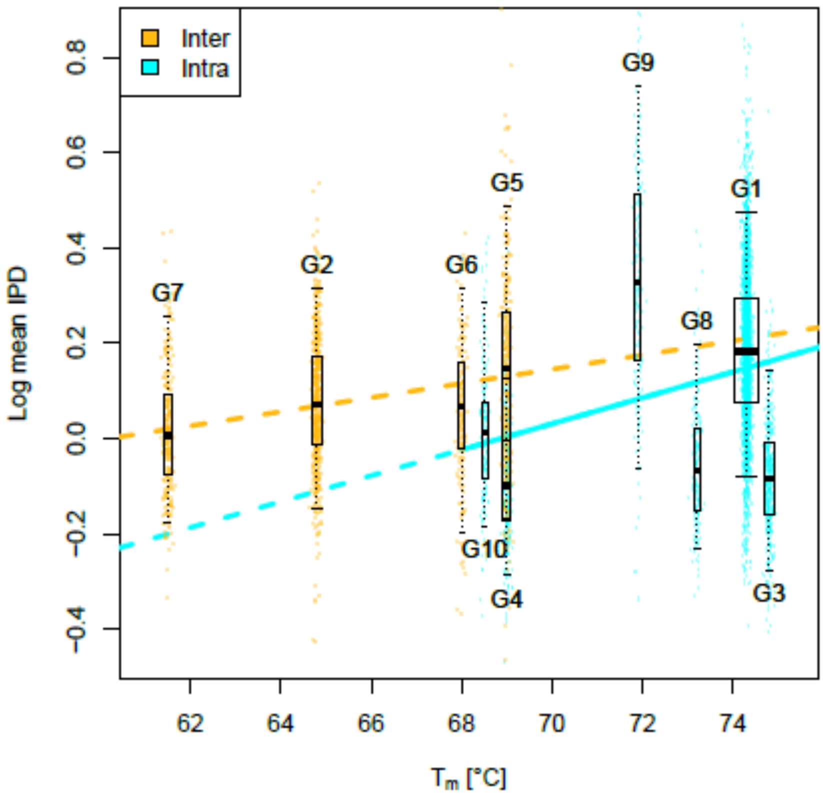
Figure S15. G-quadruplex thermostability and molecularity as predictors of polymerization kinetics.

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** 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 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.

A



B



Inter
Intra
Log mean IPD
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
G9
G5
G1
G2
G6
G7
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.

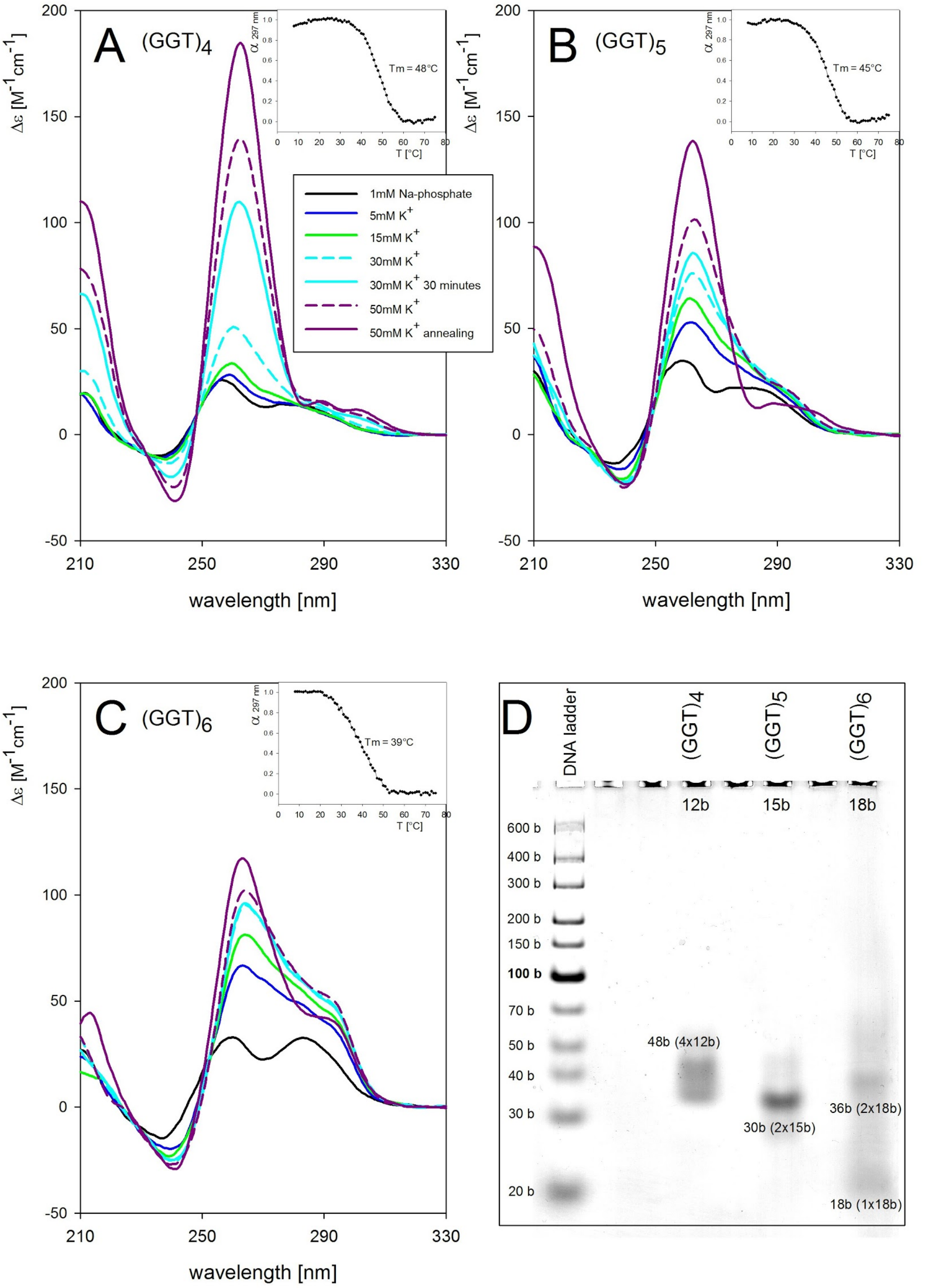
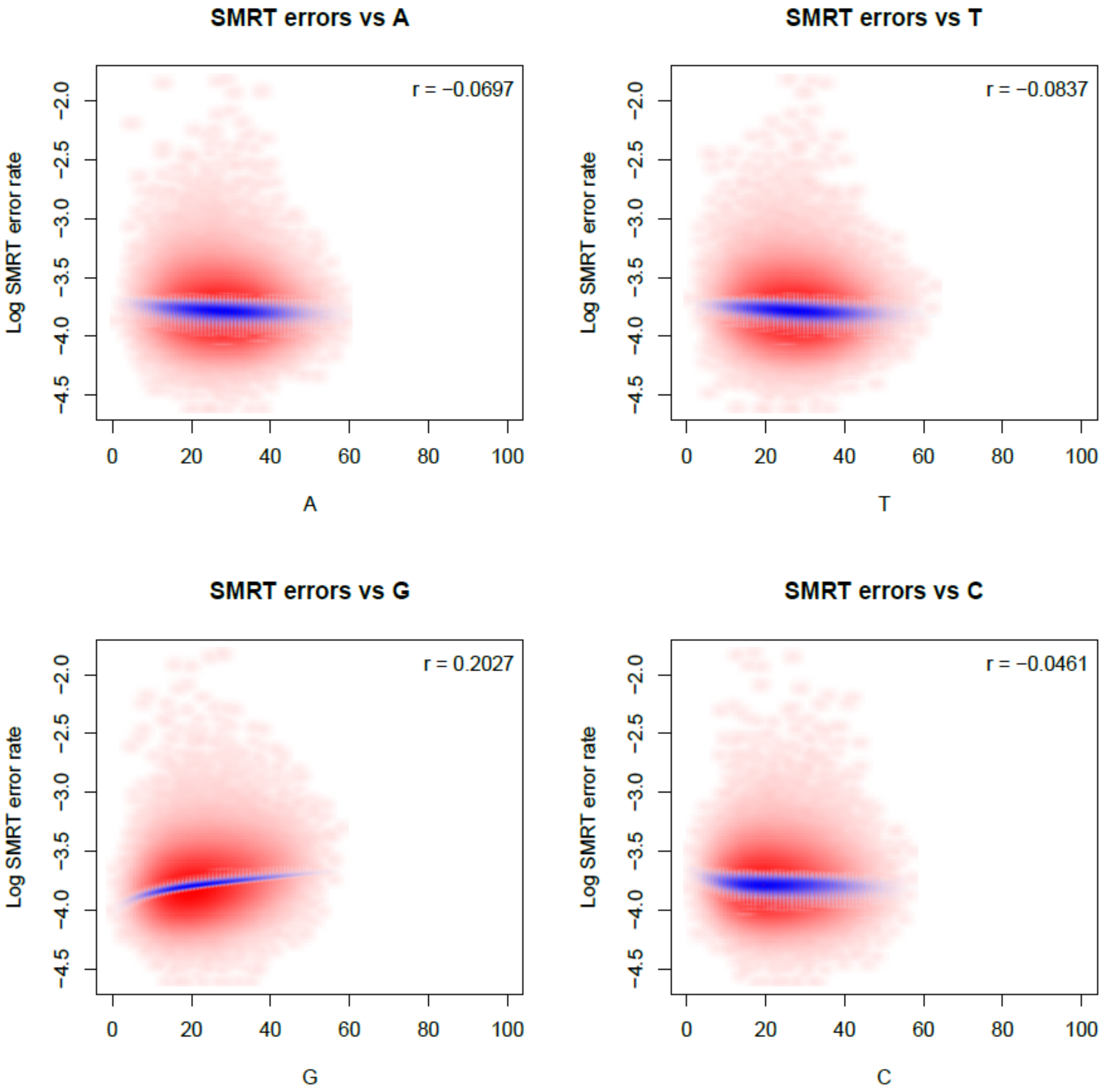


Figure S17. The relationship between error rate and sequence composition.

A. Plot of SMRT error rate 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 SMRT mismatch rates, while the blue clouds correspond to the compositional regression model with SMRT mismatch rate 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 SMRT error rate in motif-free windows. **B** Comparison between SMRT mismatch rate observed, and the one predicted using a compositional regression model fitted considering mononucleotide sequence composition on motif-free windows. Bonferroni-corrected *t*-test p-values for differences: ≤0.0001 '****', ≤0.001 '***', ≤0.01 '**', ≤0.05 '*'. Black: non-significant (corrected p-value > 0.05); red/blue: significant, with observed error rates higher/lower than composition-based predictions. Boxplot whiskers: 5th and 95th quantiles of the differences.

A.



B.

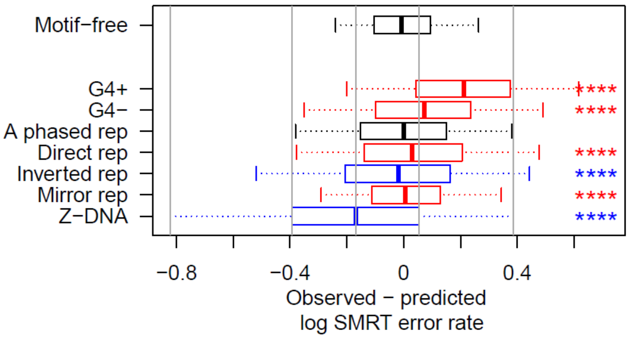
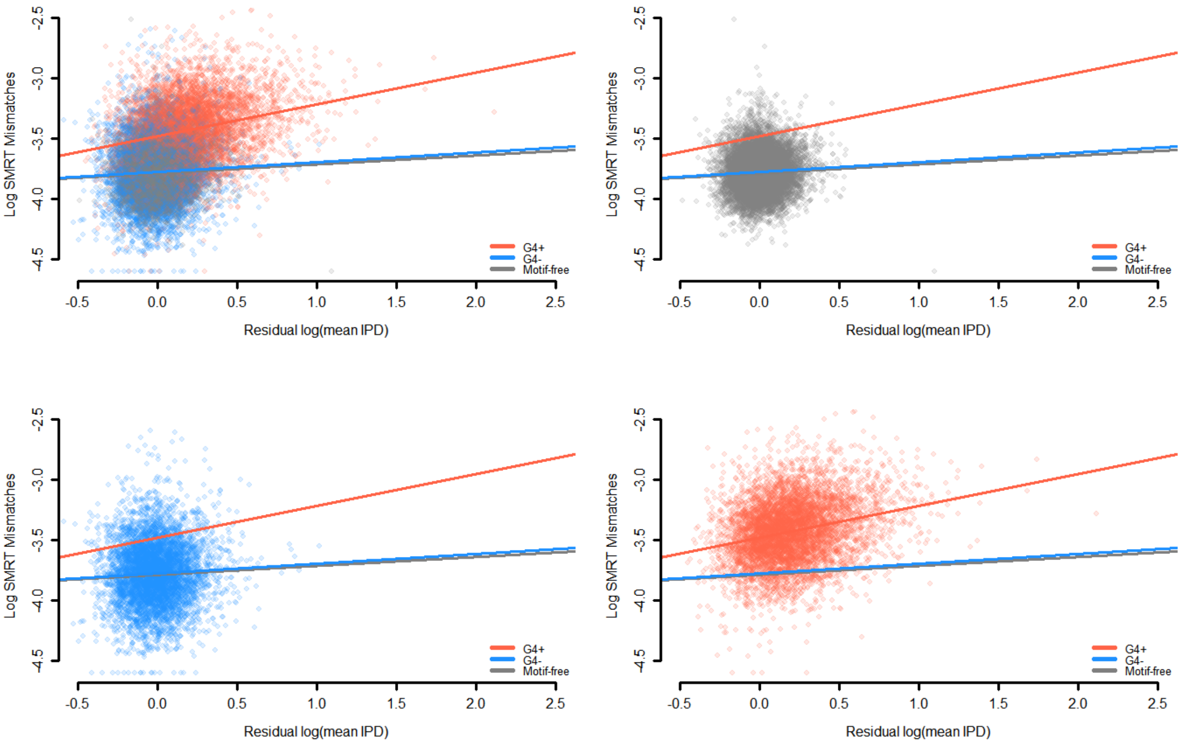

Motif−free
G4+
G4−
A phased rep
Direct rep
Inverted rep
Mirror rep
Z−DNA
****
****
****
****
****
****
−0.8
−0.4
0
0.4
Observed − predicted
log SMRT error rate

Figure S18. Errors are linked to kinetic variation. Breakdown of Figure 4. Top-left panel is a reproduction of Figure 4; top-right panel is the scatterplot of motif-free windows only; bottom-left is the scatterplot of G4- only; bottom-right is the scatterplot of G4+ only.



REFERENCES

Campos-Sánchez R, Cremona MA, Pini A, Chiaromonte F, Makova KD. 2016. Integration and Fixation Preferences of Human and Mouse Endogenous Retroviruses Uncovered with Functional Data Analysis. *PLoS Comput Biol* **12**: e1004956.

Cer RZ, Donohue DE, Mudunuri US, Temiz NA, Loss MA, Starner NJ, Halusa GN, Volfovsky N, Yi M, Luke BT, et al. 2013. Non-B DB v2.0: a database of predicted non-B DNA-forming motifs and its associated tools. *Nucleic Acids Res* **41**: D94–D100.

Cremona MA, Pini A, Cumbo F, Makova KD, Chiaromonte F, Vantini S. 2018. IWTomics: testing high-resolution sequence-based “Omics” data at multiple locations and scales. *Bioinformatics*. http://dx.doi.org/10.1093/bioinformatics/bty090.

Fungtammasan A, Ananda G, Hile SE, Su MS-W, Sun C, Harris R, Medvedev P, Eckert K, Makova KD. 2015. Accurate typing of short tandem repeats from genome-wide sequencing data and its applications. *Genome Res* **25**: 736–749.

Huertas D, Azorín F. 1996. 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.

Mirkin EV, Mirkin SM. 2007. Replication fork stalling at natural impediments. *Microbiol Mol Biol Rev* **71**: 13–35.

Pini A, Vantini S. 2017. Interval-wise testing for functional data. *J Nonparametr Stat*.

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

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

Schirmer M, D’Amore R, Ijaz UZ, Hall N, Quince C. 2016. Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. *BMC Bioinformatics* **17**: 125.

Sinden RR. 2012. *DNA Structure and Function*. Elsevier.

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

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

Yang F, Carter NP, Shi L, Ferguson-Smith MA. 1995. A comparative study of karyotypes of muntjacs by chromosome painting. *Chromosoma* **103**: 642–652.

Zhang K, Martiny AC, Reppas NB, Barry KW, Malek J, Chisholm SW, Church GM. 2006. Sequencing genomes from single cells by polymerase cloning. *Nat Biotechnol* **24**: 680–686.