Supplemental Materials

Table S1. Counts of non-B DNA loci in the human reference genome hg19 used in the study. Abbreviations in parentheses indicate feature names in coding scripts.

Non-B DNA type	Number of loci			
	Annotated genome-wide	In NCNR (used in the large-scale variation analysis)	Shorter or equal to 100 bp*	With non-overlapping flanking sequences (used in the small-scale variation analysis)*
G4 loci (WeightedScore)	670,076	178,312	177,783	50,345**
(Direct repeats (DirectCov)	1,401,209	171,653	163,021	101,199
Inverted repeats (InvertCov)	5,978,435	2,017,399	1,988,336	141,316
Mirror repeats (MirrorCov)	1,766,608	213,899	181,156	117,435
A-phased loci (APhasedCov)	375,876	146,174	137,032	99,583
Z-DNA loci (ZCov)	412,600	108,279	89,303	67,056

* A subset of loci from the previous column

**A total of 20,156 stable and 30,212 unstable loci were analyzed, as their overlaps were considered separately from all G4 loci

Table S2. Genomic landscape features analyzed in the study. Abbreviations in parentheses indicate the feature names in coding scripts.

Feature	Measurement	Present in the final model (after clustering)	Source
DNase Hypersensitive Sites (DHS)	Coverage	Yes	ENCODE (ENCODE Project Consortium 2012)
RNA polymerase II binding sites (RNA_pol_II)	Coverage	Yes	(Barski et al. 2007)
CTCF motifs (CTCF)	Signal	No	ENCODE (ENCODE Project Consortium 2012)
H2A Histone Family Member Z (H2AFZ)	Signal	Yes	ENCODE (ENCODE Project Consortium 2012)
Lamina Associated Domains (LADs)	Coverage	Yes	UCSC genome browser (Haeussler et al. 2019)
Histone 3 K27 acetylation (H3K27ac)	Signal	No	ENCODE (ENCODE Project Consortium 2012)
Histone 4 K20 methylation (H4K20me1)	Signal	No	ENCODE (ENCODE Project Consortium 2012)
Histone 3 K36 tri-methylation (H3K36me3)	Signal	Yes	ENCODE (ENCODE Project Consortium 2012)
Histone 3 K4 methylation (H3K4me1)	Signal	No	ENCODE (ENCODE Project Consortium 2012)
Histone 3 K4 di-methylation (H3K4me2)	Signal	No	ENCODE (ENCODE Project Consortium 2012)
Histone 3 K4 tri-methylation (H3K4me3)	Signal	No	ENCODE (ENCODE Project Consortium 2012)
Histone 3 K79 di-methylation (H3K79me2)	Signal	No	ENCODE (ENCODE Project Consortium 2012)
Histone 3 K9 acetylation (H3K9ac)	Signal	Yes	ENCODE (ENCODE Project Consortium 2012)
Histone 3 K9 tri-methylation (H3K9me3)	Signal	Yes	ENCODE (ENCODE Project Consortium 2012)
Histone 3 K27 tri-methylation (H3K27me3)	Signal	Yes	ENCODE (ENCODE Project Consortium 2012)
CHH methylation (chh_meth)	Coverage	Yes	(Lister et al. 2009)
CHG methylation (chg_meth)	Coverage	No	(Lister et al. 2009)
CPG methylation (cpg_meth)	Coverage	No	(Lister et al. 2009)
CpG Islands (CpGIsland)	Coverage	Yes	UCSC genome browser (Haeussler et al. 2019)
GC content (GCcontent)	Composition	No	(code available on the GitHub)
Distance to telomere (TeloDist)	Distance	Yes	(code available on the GitHub)
Distance to centromere (CentroDist)	Distance	Yes	(code available on the GitHub)
Telomere hexamer (Telo_hexamere)	Coverage	Yes	(Plohl et al. 2002)
Replication Origins (RepOrigin)	Coverage	Yes	(Besnard et al. 2012)
Replication Timing (RepTiming)	Signal	Yes	(Ryba et al. 2010)

Sex-averaged recombination rate (RecombRate)	Signal	Yes	(Kong et al. 2010)
Mappability	Signal	Yes	(Derrien et al. 2012)

Table S3. Estimated regression models for regional variation in nucleotide substitution frequencies in 1-Mb windows. See Tables S1 and S3 for abbreviations of the variable names. Variables in log scale were transformed as $\log(x+10^{-6})$ except for CpGIsland that was transformed as $\log(x+10^{-3})$. Telo_Hexamer was binarized; APhasedCov, RepTiming, RecombRate, and CHH_Meth were scaled (multiplied by 10^3, divided by 10^4, divided by 10^6 and multiplied by 10^3, respectively; see Materials and Methods). Stars indicate two-sided *t*-test significance: "***" - *p*-value<0.001; "**" - *p*-value<0.01; "*" - *p*-value<0.05; "." - *p*-value<0.1.

	SNP frequency			Fixed nucleotide substitution frequency		
	Log(Diversity)			Divergence		
	Estimate	*p*-value		Estimate	*p*-value	
(Intercept)	1.794	0.012	*	0.087	4.03E-09	***
Log WeightedScore#	-0.143	2.20E-04	***	-0.004	1.93E-06	***
(Log WeightedScore#)^2	0.008	4.55E-05	***	0.000	2.92E-06	***
APhasedCov	-0.029	0.146		-0.001	0.001	**
(APhasedCov)^2	0.003	0.171		0.000	0.001	**
InvertCov	-20.820	2.36E-04	***	-0.055	0.001	**
InvertCov^2	231.600	1.45E-04	***			
Log DirectCov	0.686	3.56E-26	***	0.001	2.31E-09	***
(Log DirectCov)^2	0.061	1.02E-20	***			
Log ZCov	0.578	0.006	**	0.022	5.55E-07	***
(Log ZCov)^2	0.040	0.009	**	0.001	2.13E-06	***
(Log MirrorCov)^2	0.004	0.016	*	0.000	0.001	**
RepTiming	-0.005	6.10E-04	***	0.000	1.36E-11	***
RecombRate	0.043	2.98E-05	***	0.000	0.400	
(RecombRate)^2	0.001	0.664		0.000	0.529	
DHS	-3.001	1.20E-06	***	-0.021	0.103	
Log RNA_pol_II	0.097	0.199				
(Log RNA_pol_II)^2	0.017	0.112		0.000	0.001	**
CHH_Meth	-0.023	0.704		0.003	0.012	*
Log CpGIsland	0.013	0.764				
(Log CpGIsland)^2	0.002	0.522		0.000	0.000	***

LADs	0.025	7.72E-05	***	0.001	5.34E-08	***
TeloDist	-1.112	3.12E-43	***	-0.028	2.30E-59	***
TeloDist^2	1.454	3.64E-24	***	0.041	2.02E-39	***
CentroDist	-0.082	0.077	.	-0.003	0.010	*
CentroDist^2	-0.072	0.272		0.002	0.163	
H2AFZ				0.000	0.012	*
H2AFZ^2	0.001	0.004	**			
Log H3K27me3	-0.017	0.078	.	0.000	0.855	
Log H3K36me3	-0.056	0.111		-0.002	0.004	**
Log H3K9ac	-0.064	0.419		0.001	0.074	.
(Log H3K9ac)^2	0.085	0.068	.			
Log H3K9me3	0.097	8.51E-09	***	0.011	1.60E-08	***
(Log H3K9me3)^2				-0.004	1.99E-09	***
Telo_Hexamere	-0.004	0.548		0.000	0.013	*
Mappability	-0.778	8.96E-10	***	0.042	9.54E-52	***

[#]Coverage of G4 loci weighted by their stability

Table S4. Frequencies of polymorphic substitutions corrected by their trinucleotide context. We used Fisher's exact test to evaluate whether the frequency of each substitution type at the first flanking base of non-B DNA loci was significantly different from that at control sequences. We only considered trinucleotides found adjacent to the non-B DNA loci, and not all possible trinucleotide combinations. Fisher's exact test's *p*-values were adjusted for multiple testing using Bonferroni correction. Substitution types with zero counts had no calculable values and were assigned an 'NA'. Odds ratio computed with NAs resulted in a 'nan' (not a number). Identical trinucleotides present at the immediate flanking positions of both non-B DNA loci and controls were compared, whereas trinucleotides present only in controls were discarded.

A. Stable G4 loci

Substitution type	First upstream flanking base				First downstream flanking base			
	Substitution frequency in non-B DNA	Substitution frequency in control sequences	Odds ratio	Adjusted *p*-value	Substitution frequency in non-B DNA	Substitution frequency in control sequences	Odds ratio	Adjusted *p*-value
A -> C	1.89E-03	2.34E-03	0.81	1.00E+00	2.55E-03	2.79E-03	0.91	7.50E-01
A -> G	1.61E-02	4.47E-03	3.61	1.34E-06	1.20E-02	4.53E-03	2.66	3.97E-03
A -> T	1.89E-03	1.95E-03	0.97	1.00E+00	2.74E-03	1.53E-03	1.79	7.57E-01
C -> A	1.48E-02	1.21E-02	1.22	1.00E+00	5.22E-03	3.82E-03	1.37	8.02E-01
C -> G	2.10E-02	1.04E-02	2.03	1.00E+00	4.80E-03	3.47E-03	1.38	5.47E-01
C -> T	1.22E-01	1.15E-01	1.06	1.00E+00	1.79E-02	1.39E-02	1.29	2.92E-01
G -> A	1.85E-02	9.65E-03	1.92	1.00E+00	1.87E-02	1.11E-02	1.68	2.69E-01
G -> C	9.90E-03	3.81E-03	2.60	1.00E+00	NA	NA	NA	1.00E+00
G -> T	NA	NA	nan	NA	2.90E-02	6.82E-03	4.25	3.76E-01
T -> A	1.86E-03	3.32E-03	0.56	1.00E+00	3.03E-03	2.99E-03	1.01	1.00E+00
T -> C	9.38E-03	6.60E-03	1.42	1.00E+00	4.25E-03	6.88E-03	0.62	3.20E-01
T -> G	3.53E-02	2.51E-03	14.04	1.09E-30	3.57E-02	4.20E-03	8.51	1.60E-13

B. Unstable G4 loci

	First upstream flanking base				First downstream flanking base			
Substitution type	**Substitution frequency in non-B DNA**	**Substitution frequency in control sequences**	**Odds ratio**	**Adjusted *p*-value**	**Substitution frequency in non-B DNA**	**Substitution frequency in control sequences**	**Odds ratio**	**Adjusted *p*-value**
A -> C	1.82E-03	2.34E-03	0.78	1.00E+00	9.24E-04	2.79E-03	0.33	1.00E+00
A -> G	6.52E-03	4.47E-03	1.46	1.00E+00	5.69E-03	4.53E-03	1.26	1.00E+00
A -> T	1.09E-03	1.95E-03	0.56	1.00E+00	1.23E-03	1.53E-03	0.80	1.00E+00
C -> A	1.06E-02	1.21E-02	0.87	1.00E+00	3.77E-03	3.82E-03	0.99	1.00E+00
C -> G	1.22E-02	1.04E-02	1.18	1.00E+00	2.55E-03	3.47E-03	0.74	1.00E+00
C -> T	1.19E-01	1.15E-01	1.04	1.00E+00	1.50E-02	1.39E-02	1.08	1.00E+00
G -> A	1.74E-02	1.21E-02	1.43	1.00E+00	1.42E-02	1.11E-02	1.27	1.00E+00
G -> C	1.10E-02	3.26E-03	3.39	2.88E-01	7.19E-03	4.10E-03	1.76	1.00E+00
G -> T	9.46E-03	2.33E-03	4.07	2.69E-01	1.17E-02	7.47E-03	1.57	1.00E+00
T -> A	2.05E-03	3.32E-03	0.62	1.00E+00	1.73E-03	3.00E-03	0.58	1.00E+00
T -> C	1.10E-02	6.60E-03	1.66	3.52E-01	3.87E-03	6.88E-03	0.56	1.00E+00
T -> G	2.73E-03	2.51E-03	1.08	1.00E+00	8.66E-03	4.20E-03	2.06	9.37E-01

C. Inverted repeats

	First upstream flanking base				First downstream flanking base			
Substitution type	**Substitution frequency in non-B DNA**	**Substitution frequency in control sequences**	**Odds ratio**	**Adjusted *p*-value**	**Substitution frequency in non-B DNA**	**Substitution frequency in control sequences**	**Odds ratio**	**Adjusted *p*-value**
A -> C	2.15E-03	1.69E-03	1.27	1.00E+00	1.73E-03	1.80E-03	0.96	1.00E+00
A -> G	6.41E-03	6.66E-03	0.96	1.00E+00	6.34E-03	6.16E-03	1.03	1.00E+00
A -> T	2.85E-03	1.8E-03	1.55	4.5E-02	2.54E-03	1.40E-03	1.81	4.39E-03
C -> A	4.31E-03	3.79E-03	1.14	1.00E+00	4.53E-03	3.47E-03	1.31	6.33E-01
C -> G	2.72E-03	3.41E-03	0.80	1.00E+00	2.85E-03	3.00E-03	0.95	1.00E+00
C -> T	1.43E-02	1.36E-02	1.05	1.00E+00	1.36E-02	1.27E-02	1.07	1.00E+00
G -> A	1.42E-02	1.39E-02	1.03	1.00E+00	1.42E-02	1.207E-02	1.19	2.18E-01
G -> C	3.29E-03	3.32E-03	0.99	1.00E+00	2.74E-03	2.76E-03	0.99	1.00E+00
G -> T	4.13E-03	3.58E-03	1.16	1.00E+00	3.85E-03	3.17E-03	1.21	1.00E+00
T -> A	2.26E-03	2.11E-03	1.07	1.00E+00	4.13E-03	2.10E-03	1.96	9.00E-06
T -> C	6.40E-03	6.78E-03	0.95	1.00E+00	7.78E-03	6.29E-03	1.24	1.33E-01
T -> G	1.83E-03	2.04E-03	0.90	1.00E+00	2.36E-03	2.04E-03	1.16	1.00E+00

D. Direct repeats

	First upstream flanking base				First downstream flanking base			
Substitution type	**Substitution frequency in non-B DNA**	**Substitution frequency in control sequences**	**Odds ratio**	**Adjusted *p*-value**	**Substitution frequency in non-B DNA**	**Substitution frequency in control sequences**	**Odds ratio**	**Adjusted *p*-value**
A -> C	5.92E-03	2.07E-03	2.87	2.76E-13	5.47E-03	1.95E-03	2.81	8.52E-13
A -> G	2.01E-02	6.80E-03	2.96	9.07E-48	1.66E-02	6.76E-03	2.45	2.02E-30
A -> T	5.35E-03	1.66E-03	3.23	1.35E-14	5.23E-03	1.83E-03	2.86	1.62E-12
C -> A	1.01E-02	3.78E-03	2.68	3.95E-13	7.02E-03	3.46E-03	2.03	3.33E-06
C -> G	7.60E-03	3.86E-03	1.97	3.92E-06	8.85E-03	3.19E-03	2.78	1.20E-13
C -> T	2.74E-02	1.42E-02	1.94	9.38E-21	2.87E-02	1.52E-02	1.89	1.61E-20
G -> A	2.69E-02	1.37E-02	1.97	2.78E-21	2.60E-02	1.22E-02	2.13	3.22E-24
G -> C	7.36E-03	2.73E-03	2.69	1.22E-10	7.39E-03	3.35E-03	2.21	1.33E-07
G -> T	8.59E-03	3.23E-03	2.66	1.52E-12	9.03E-03	3.25E-03	2.78	1.43E-13
T -> A	4.66E-03	2.00E-03	2.34	1.37E-07	5.32E-03	2.08E-03	2.56	1.56E-09
T -> C	1.64E-02	6.52E-03	2.52	9.76E-32	1.77E-02	7.16E-03	2.47	4.99E-32
T -> G	6.00E-03	1.79E-03	3.35	1.09E-16	5.91E-03	2.41E-03	2.45	6.64E-11

E. Mirror repeats

	First upstream flanking base				First downstream flanking base			
Substitution type	**Substitution frequency in non-B DNA**	**Substitution frequency in control sequences**	**Odds ratio**	**Adjusted *p*-value**	**Substitution frequency in non-B DNA**	**Substitution frequency in control sequences**	**Odds ratio**	**Adjusted *p*-value**
A -> C	2.97E-03	1.92E-03	1.55	1.23E-01	3.05E-03	2.07E-03	1.47	1.64E-01
A -> G	8.46E-03	6.41E-03	1.32	2.08E-02	7.76E-03	6.71E-03	1.16	1.00E+00
A -> T	2.73E-03	2.04E-03	1.34	8.3E-01	3.18E-03	1.70E-03	1.88	7.31E-04
C -> A	6.26E-03	3.32E-03	1.88	5.70E-05	5.68E-03	3.89E-03	1.46	6.18E-02
C -> G	3.93E-03	2.88E-03	1.36	6.62E-01	2.92E-03	2.96E-03	0.99	1.00E+00
C -> T	1.46E-02	1.35E-02	1.08	1.00E+00	1.58E-02	1.41E-02	1.13	1.00E+00
G -> A	1.63E-02	1.41E-02	1.16	6.28E-01	1.33E-02	1.55E-02	0.86	6.45E-01
G -> C	4.02E-03	2.85E-03	1.414	4.01E-01	3.40E-03	3.45E-03	0.99	1.00E+00
G -> T	5.69E-03	3.37E-03	1.69	2.58E-03	5.39E-03	3.54E-03	1.52	3.46E-02
T -> A	3.80E-03	2.17E-03	1.75	1.87E-03	2.92E-03	1.56E-03	1.87	2.07E-03
T -> C	7.70E-03	6.45E-03	1.19	6.49E-01	7.66E-03	6.15E-03	1.24	2.12E-01
T -> G	3.22E-03	1.82E-03	1.77	2.98E-03	3.18E-03	2.15E-03	1.48	1.08E-01

F. Z-DNA

	First upstream flanking base				First downstream flanking base			
Substitution type	**Substitution frequency in non-B DNA**	**Substitution frequency in control sequences**	**Odds ratio**	**Adjusted *p*-value**	**Substitution frequency in non-B DNA**	**Substitution frequency in control sequences**	**Odds ratio**	**Adjusted *p*-value**
A -> C	7.94E-03	1.45E-03	5.46	1.799E-21	1.61E-02	1.98E-03	8.15	9.01E-47
A -> G	1.84E-02	6.73E-03	2.74	7.58E-27	3.60E-02	6.52E-03	5.52	4.67E-87
A -> T	5.64E-03	1.79E-03	3.15	1.02E-08	1.08E-02	1.96E-03	5.52	1.84E-25
C -> A	1.28E-02	3.38E-03	3.77	6.09E-11	1.19E-02	3.66E-03	3.26	2.10E-08
C -> G	1.51E-02	3.40E-03	4.45	3.22E-14	8.60E-03	3.24E-03	2.65	2.55E-03
C -> T	8.15E-03	9.83E-03	0.83	1.00E+00	1.44E-02	1.45E-02	0.99	1.00E+00
G -> A	1.26E-02	1.18E-02	1.06	1.00E+00	9.13E-03	8.27E-03	1.10	1.00E+00
G -> C	8.75E-03	3.08E-03	2.83	3.07E-05	1.46E-02	2.71E-03	5.40	1.95E-17
G -> T	1.44E-02	3.23E-03	4.46	6.23E-10	1.33E-02	3.31E-03	4.02	9.71E-12
T -> A	6.07E-03	1.69E-03	3.59	1.48E-09	5.61E-03	1.97E-03	2.84	2.75E-08
T -> C	7.94E-02	1.45E-03	5.46	1.798E-21	1.87E-02	7.28E-03	2.56	1.50E-24
T -> G	1.90E-02	1.46E-03	13.0	6.87E-59	1.00E-02	1.45E-03	6.93	8.16E-32

G. A-phased repeats

	First upstream flanking base				First downstream flanking base			
Substitution type	**Substitution frequency in non-B DNA**	**Substitution frequency in control sequences**	**Odds ratio**	**Adjusted *p*-value**	**Substitution frequency in non-B DNA**	**Substitution frequency in control sequences**	**Odds ratio**	**Adjusted *p*-value**
A -> C	NA	NA	nan	1.00E+00	NA	NA	nan	1.00E+00
A -> G	NA	NA	nan	1.00E+00	NA	NA	nan	1.00E+00
A -> T	NA	NA	nan	1.00E+00	NA	NA	nan	1.00E+00
C -> A	4.56E-03	2.43E-03	1.87	1.93E-03	3.92E-03	3.64E-03	1.08	1.00E+00
C -> G	3.42E-03	2.86E-03	1.19	1.00E+00	2.62E-03	2.30E-03	1.14	2.09E-01
C -> T	8.92E-03	8.30E-03	1.07	1.00E+00	1.11E-02	1.45E-02	0.76	2.51E-02
G -> A	1.07E-02	1.43E-02	0.75	1.47E-02	8.15E-03	8.17E-03	1.00	1.00E+00
G -> C	2.36E-03	4.06E-03	0.58	1.81E-02	3.05E-03	2.37E-03	1.28	1.00E+00
G -> T	3.64E-03	4.24E-03	0.86	1.00E+00	5.10E-03	2.93E-03	1.74	2.94E-03
T -> A	NA	NA	nan	1.00E+00	NA	NA	nan	1.00E+00
T -> C	NA	NA	nan	1.00E+00	NA	NA	nan	1.00E+00
T -> G	NA	NA	nan	1.00E+00	NA	NA	nan	1.00E+00

Figure S1. Construction of Non-Coding Non-Repetitive (NCNR) subgenome. (**A**) Exons and UTRs were extended to account for variant hitchhiking potentially present in these regions. (**B**) Variants were filtered out when not intersecting with the NCNR genome.

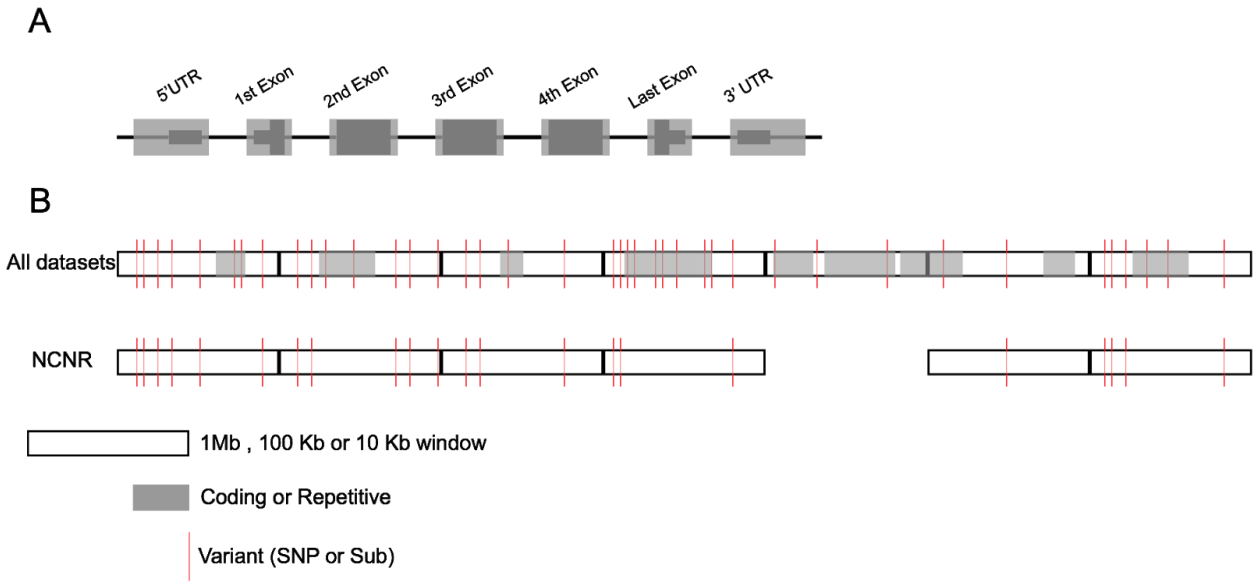


Figure S2. The procedure followed to select one non-B DNA locus from a set of overlapping loci of the same type. One locus was chosen at random, the others were discarded.

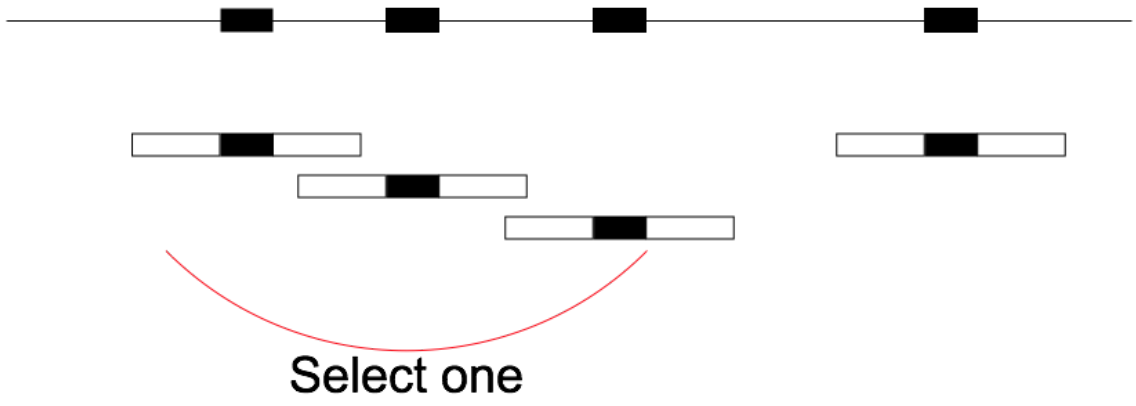


Figure S3. Scaling schema for non-B DNA loci. Because the scaled interval may have a higher resolution than the original motif interval, a SNP or FNS may occupy more than one scaled position. The scaled interval has 180 bins.

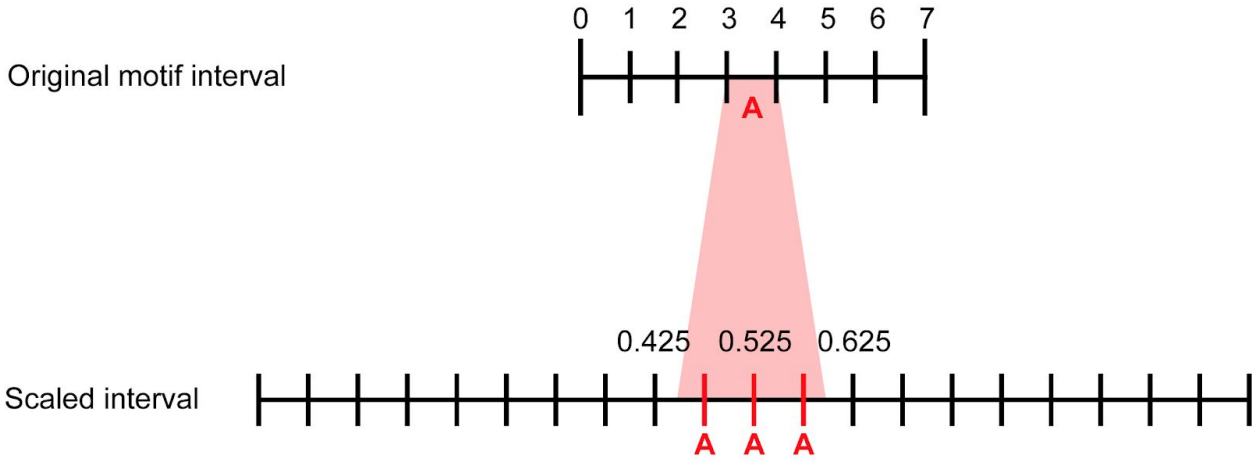
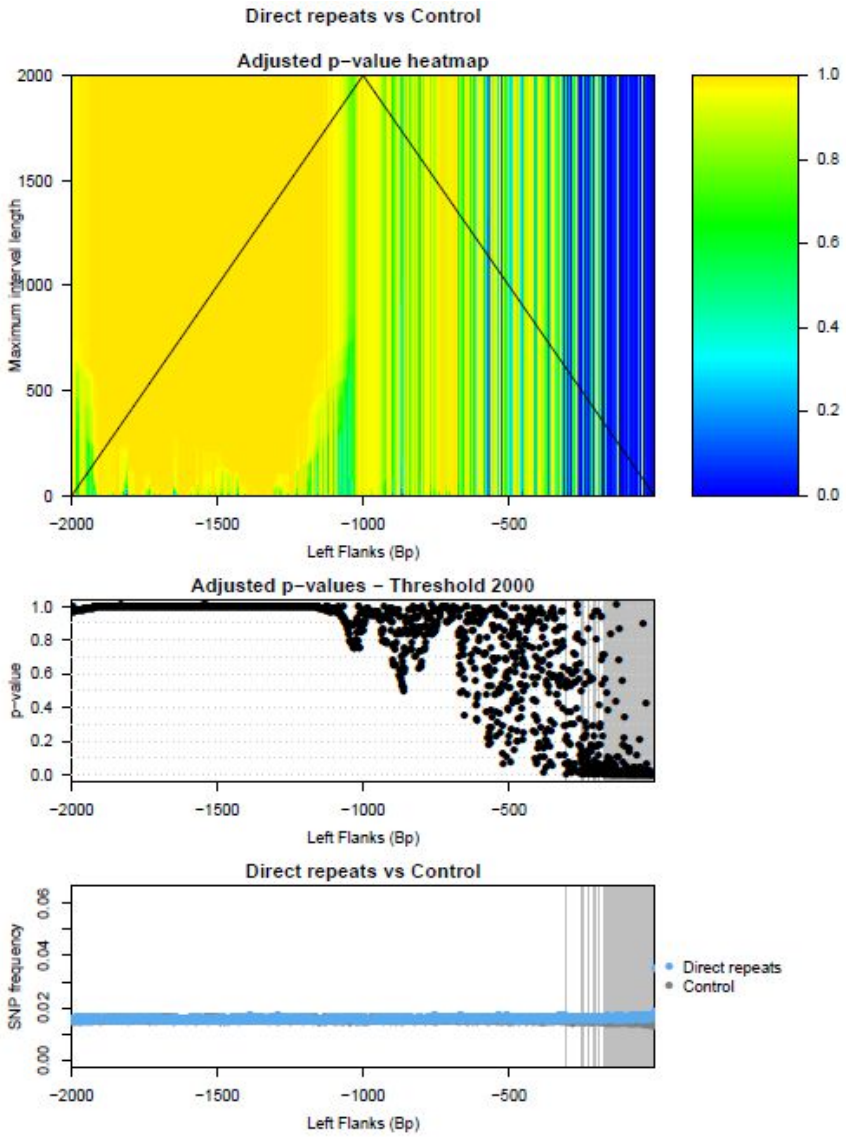
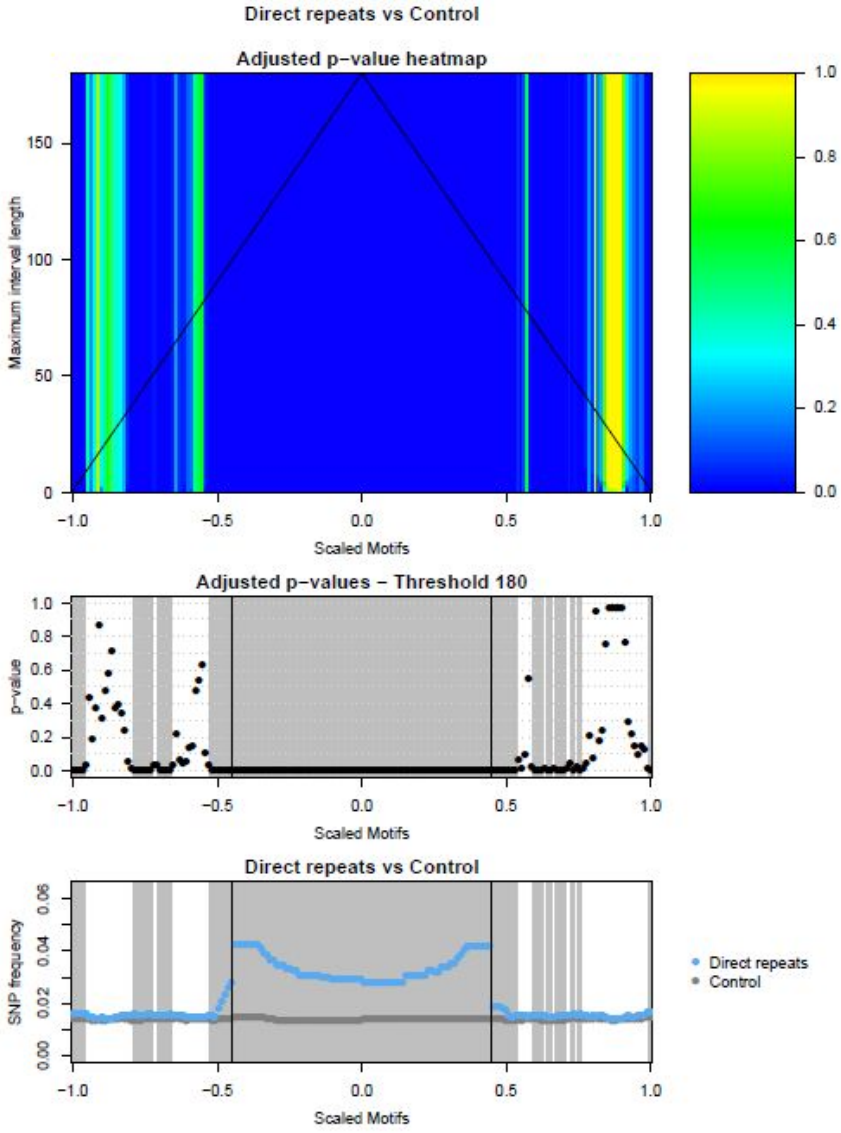


Figure S4. Example of complete IWTomics output for the comparison of single-nucleotide polymorphism (SNP) frequencies between non-B DNA and control sequences. (**A**) Flanking region upstream of the loci; (**B**) direct repeats loci; (**C**) flanking region downstream of the loci. In each figure, the top heatmap shows the adjusted *p*-value curve for each possible scale (from 1 to 2,000 bp in **A** and **C**, from 1 to 180 bins in **B**; each row corresponds to one scale). The central panel shows the adjusted *p*-value curve corresponding to the selected scale threshold (2,000 bp in **A** and **C**, 180 bins in **B**), while the bottom panel shows the mean SNP frequency curves in the two groups. Gray areas indicate positions with significant differences (adjusted *p*-value<0.01). The different sections of the non-B DNA locus are shown by vertical black lines in **B**.

A



B


Direct repeats vs Control
Adjusted p−value heatmap
Maximum interval length
150
100
50
0
−1.0
−0.5
0.0
0.5
1.0
Scaled Motifs
1.0
0.8
0.6
0.4
0.2
0.0
Adjusted p−values − Threshold 180
p−value
1.0
0.8
0.6
0.4
0.2
0.0
−1.0
−0.5
0.0
0.5
1.0
Scaled Motifs
Direct repeats vs Control
SNP frequency
0.06
0.04
0.02
0.00
−1.0
−0.5
0.0
0.5
1.0
Scaled Motifs
Direct repeats
Control

C

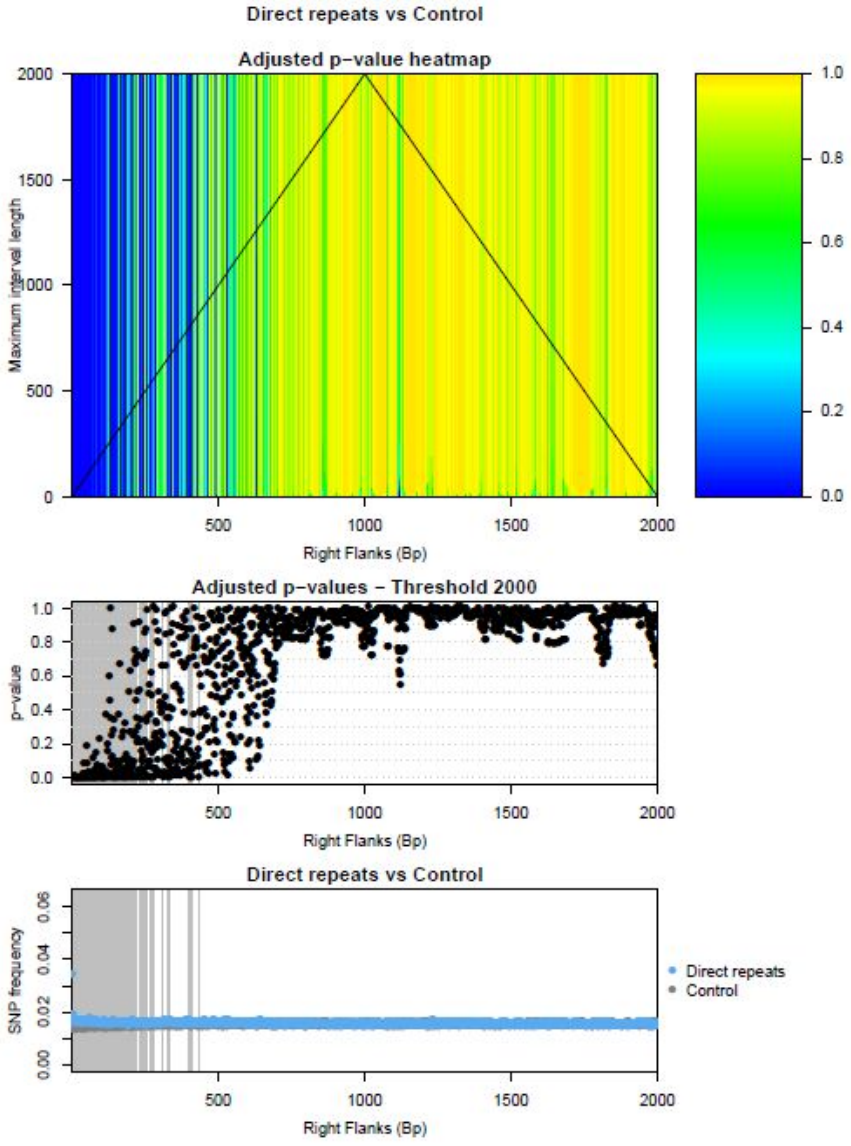
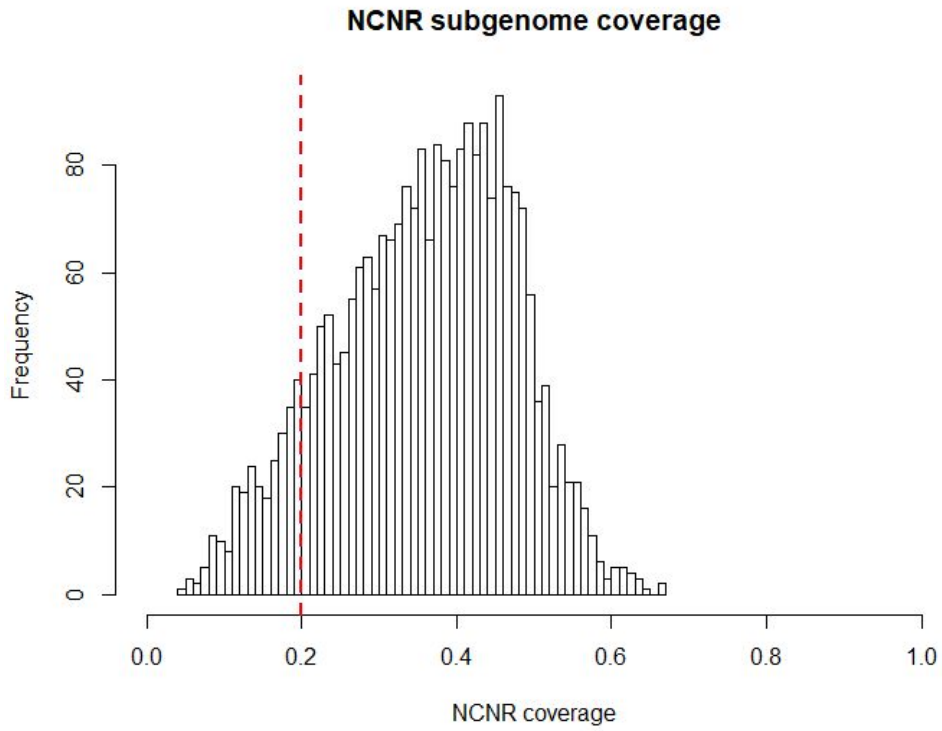

Direct repeats vs Control
Adjusted p−value heatmap
Maximum interval length
Right Flanks (Bp)
Adjusted p−values − Threshold 2000
p−value
Right Flanks (Bp)
Direct repeats vs Control
SNP frequency
Direct repeats
Control
Right Flanks (Bp)

Figure S5. Histogram of (A) NCNR subgenome coverage and (B) mappability in all 1-Mb windows. Vertical red lines show the employed thresholds.

A



B

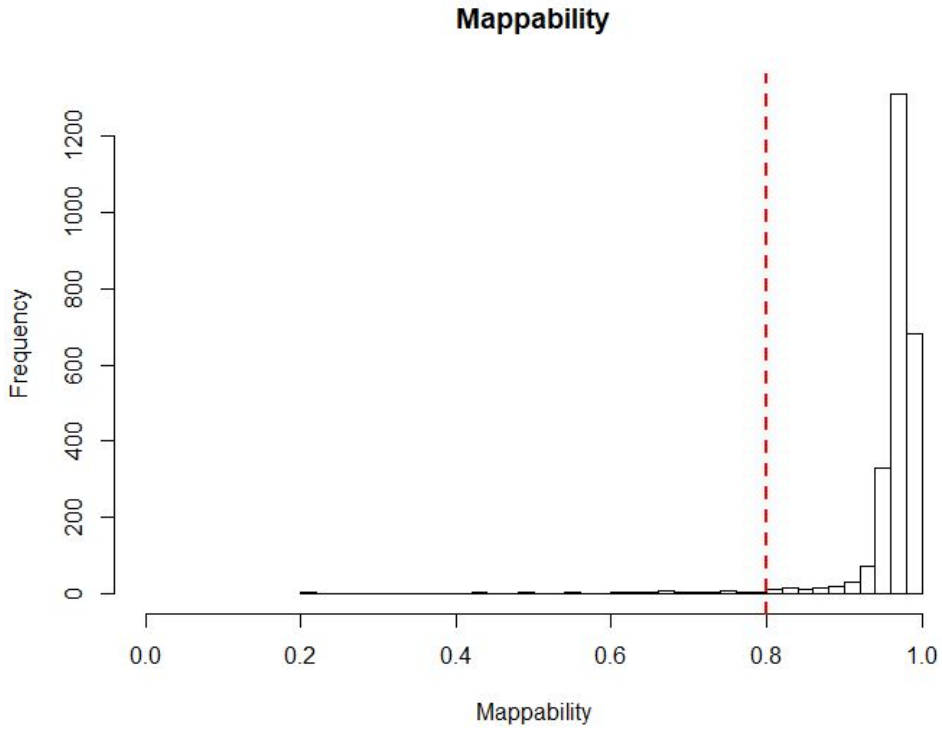
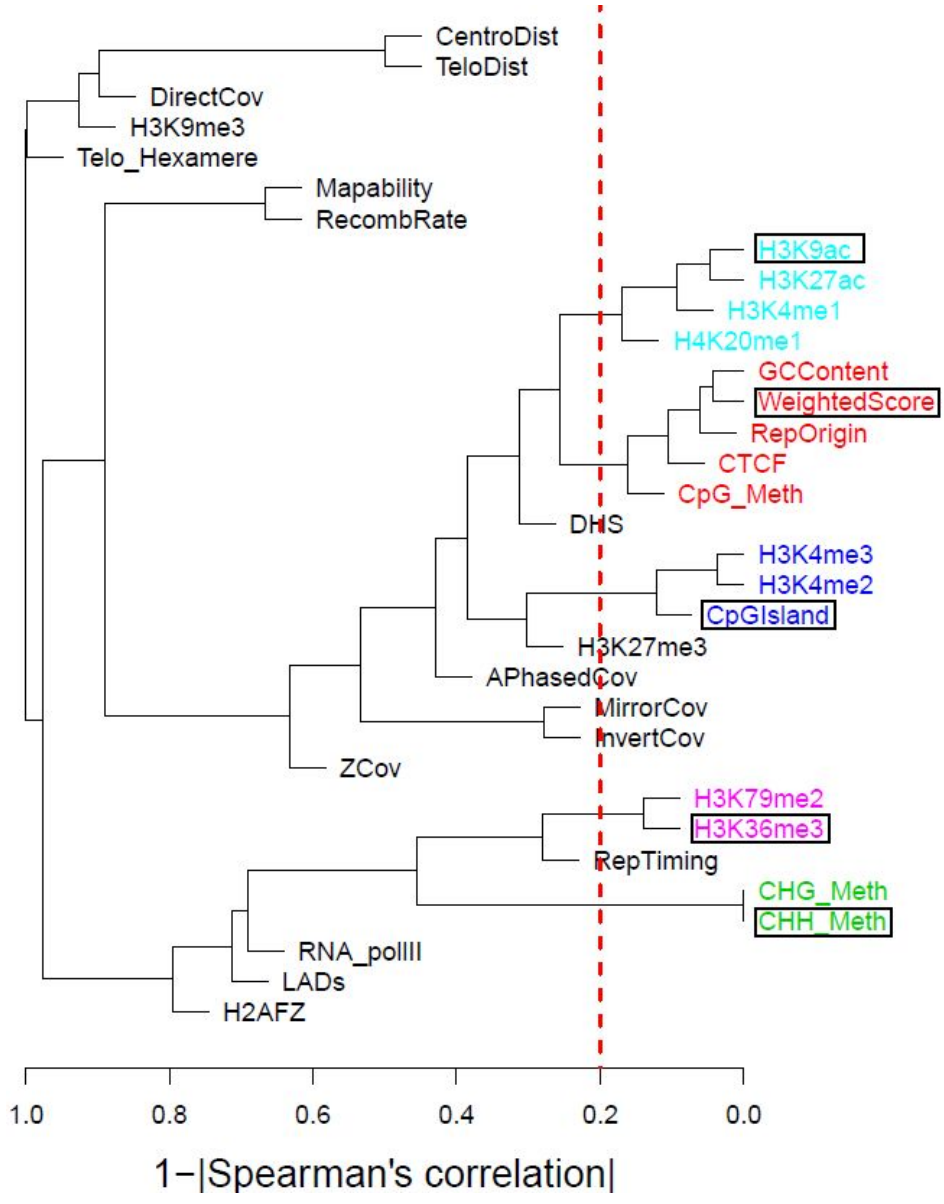


Figure S6. Predictor clustering by (A) Spearman's and (B) Pearson's correlations. Only one variable per cluster, defined using a threshold of 0.8 on the absolute value of the correlation coefficient, was selected for the regression model. Variables in the same clusters are shown using the same color. Rectangles indicate the selected variable in each cluster. 'WeightedScore' is the Quadron score weighted average of the per-window coverage of G-quadruplexes. See Tables S1 and S3 for the abbreviations of variable names.

A



B

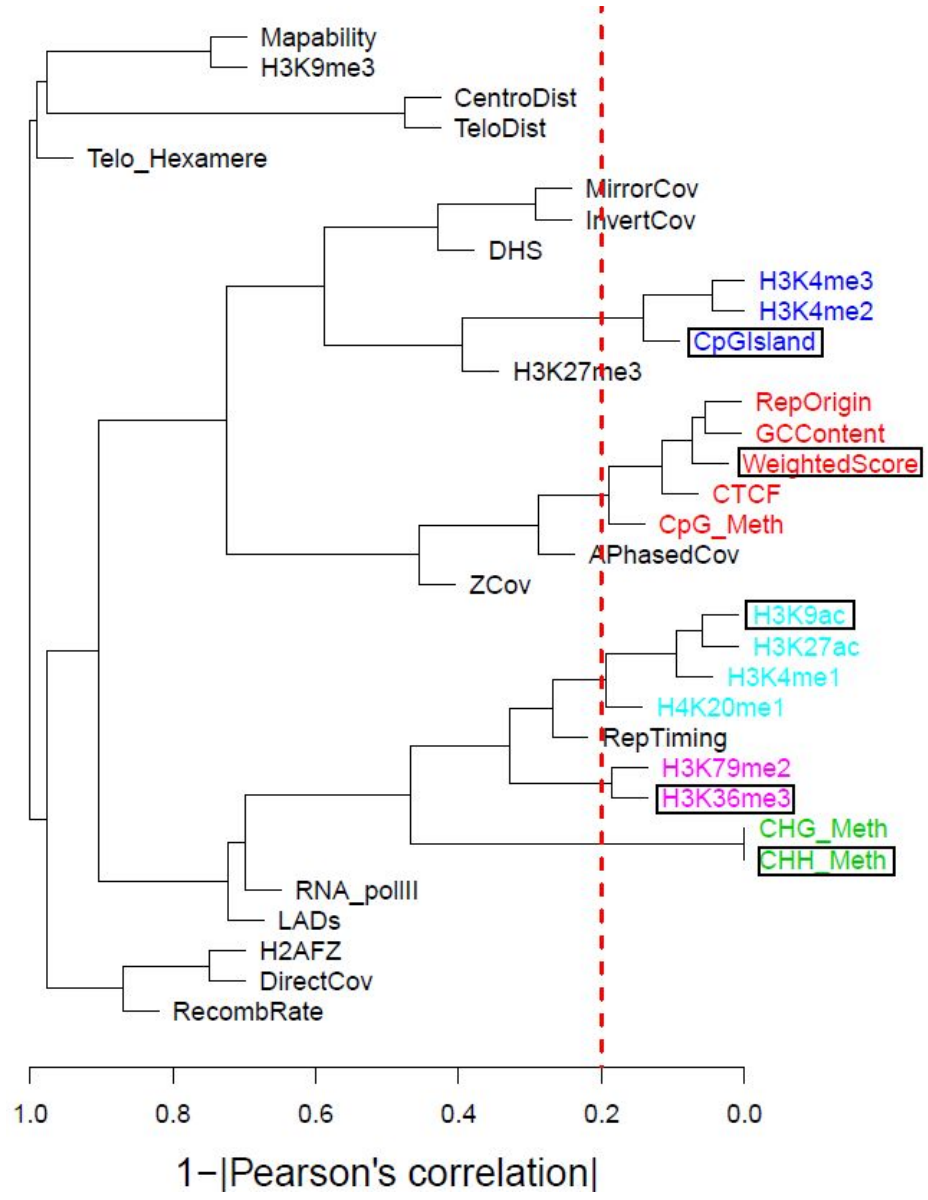
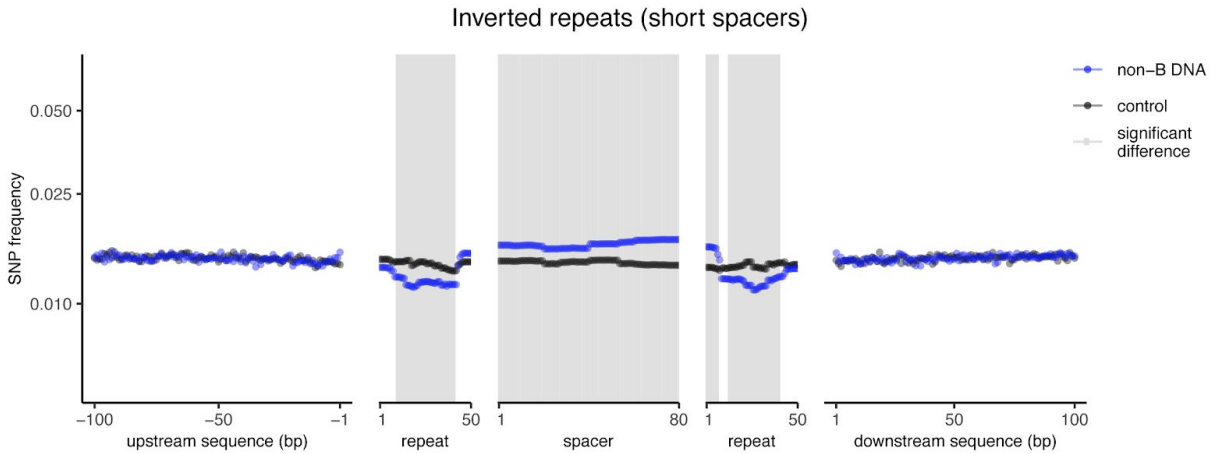
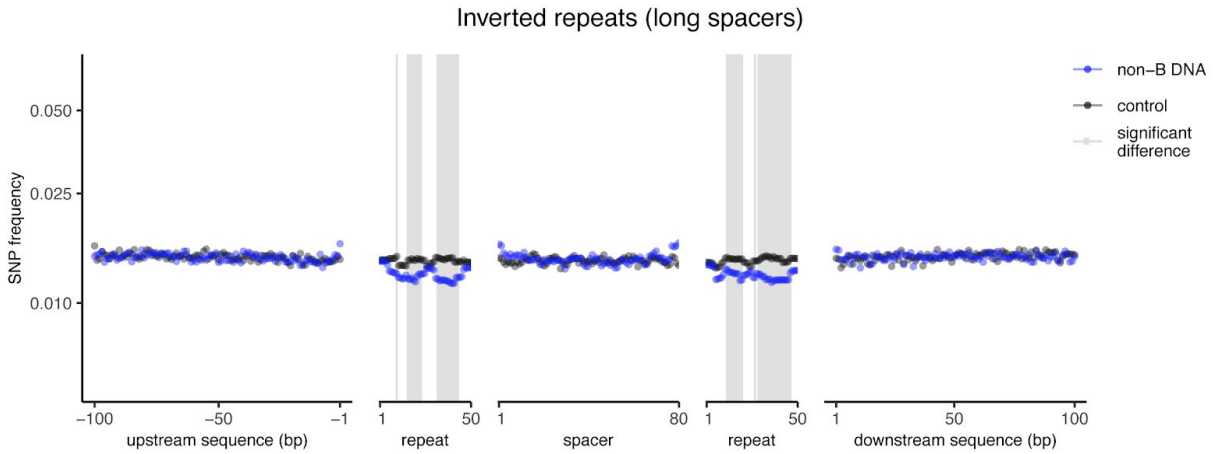

Mapability
H3K9me3
CentroDist
TeloDist
Telo_Hexamere
MirrorCov
InvertCov
DHS
H3K4me3
H3K4me2
CpGIsland
H3K27me3
RepOrigin
GCContent
WeightedScore
CTCF
CpG_Meth
APhasedCov
ZCov
H3K9ac
H3K27ac
H3K4me1
H4K20me1
RepTiming
H3K79me2
H3K36me3
CHG_Meth
CHH_Meth
RNA_polIII
LADs
H2AFZ
DirectCov
RecombRate
1.0
0.8
0.6
0.4
0.2
0.0
1−|Pearson's correlation|

Figure S7. Genome-wide nucleotide substitution frequencies at inverted and mirror repeats considering different lengths of spacers. Inverted and mirror repeats are broken into spacers and repeat arms. The positions of nucleotide substitutions within motifs were scaled based on motif size (see Materials and Methods for details). Gray areas indicate significantly different rates between groups (IWTomics adjusted *p*-value curve <0.01). Single-nucleotide substitution (SNP) and fixed nucleotide substitution (FNS) frequencies for inverted repeats vs. control sequences (**A, B**), mirror repeats vs. control sequences (**C, D**) with <15-bp spacers (**A, C**), ≥15-bp spacers (**B, D**).

A



B



C


Mirror repeats (short spacers)
non−B DNA
control
significant difference
SNP frequency
0.050
0.025
0.010
−100
−50
−1
upstream sequence (bp)
1
50
repeat
1
80
spacer
1
50
repeat
1
50
100
downstream sequence (bp)

D

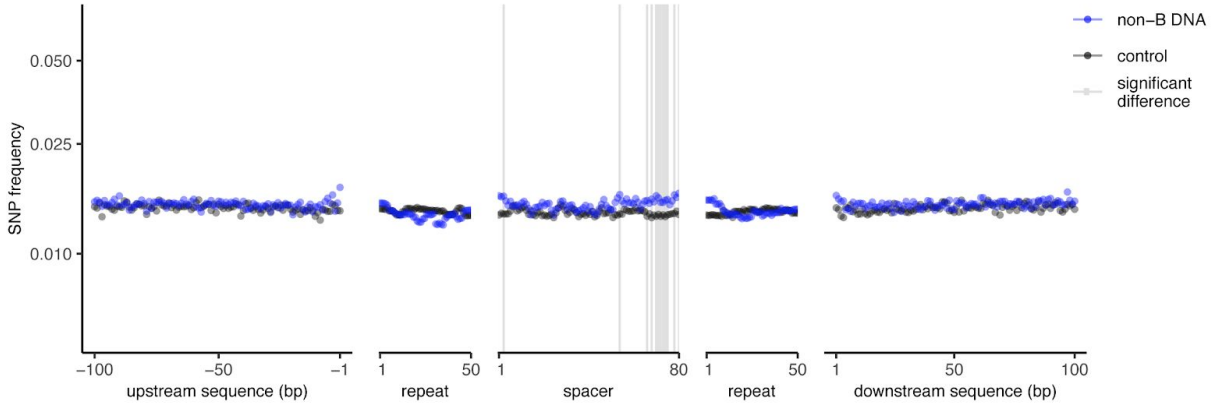
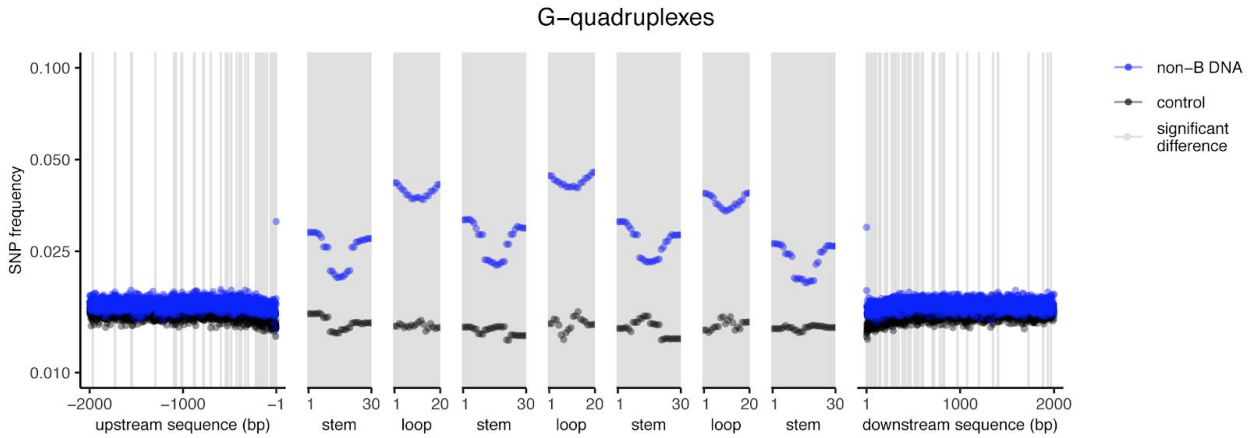
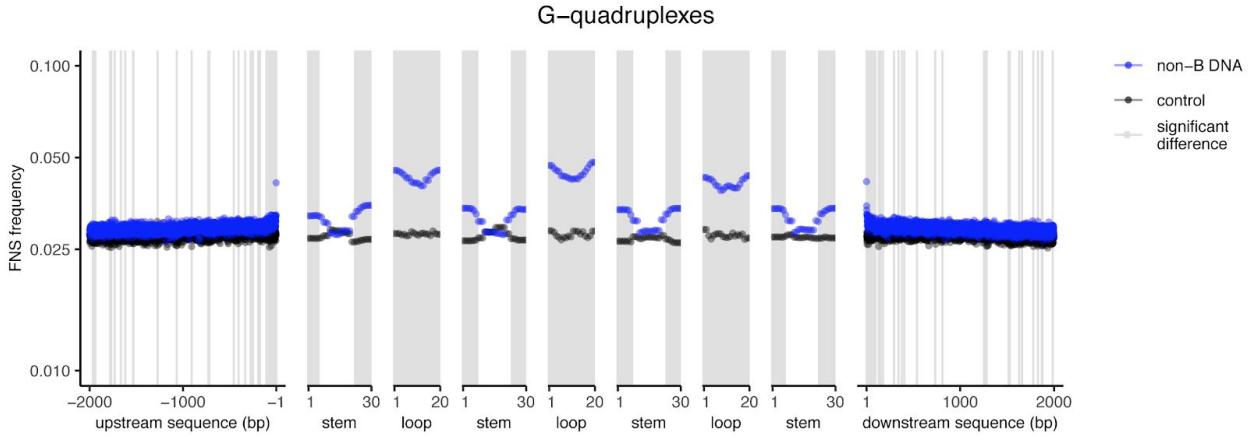

Mirror repeats (long spacers)
non−B DNA
control
significant difference
SNP frequency
0.050
0.025
0.010
−100
−50
−1
upstream sequence (bp)
1
50
repeat
1
80
spacer
1
50
repeat
1
50
100
downstream sequence (bp)

Figure S8. Genome-wide nucleotide substitution frequencies at non-B loci with the whole 2-kb flanking regions shown. The positions of nucleotide substitutions within motifs were scaled based on motif size (see Materials and Methods for details). Flanking regions are the 2 kilobases up- and downstream from the loci. Y-axes are shown on a log scale for better visualization. Gray areas indicate significantly different rates between groups (IWTomics adjusted *p*-value curve <0.01). Single-nucleotide substitution (SNP) and fixed nucleotide substitution (FNS) frequencies for G-quadruplexes vs. control sequences (**A, B**), stable vs. unstable G-quadruplexes (**C, D**), stable G-quadruplexes vs. control sequences (**E, F**), unstable G-quadruplexes vs. control sequences (**G, H**), inverted repeats vs. control sequences (**I, J**), direct repeats vs. control sequences (**K, L**), mirror repeats vs. control sequences (**M, N**), Z-DNA vs. control sequences (**O, P**), and A-phased repeats (**Q, R**). For G-quadruplexes, stems are runs of guanines and loops are unspecified nucleotides between stems. Inverted, direct, and mirror repeats are broken into spacers and repeat arms, and A-phased repeats are broken into A-tracts and spacers.

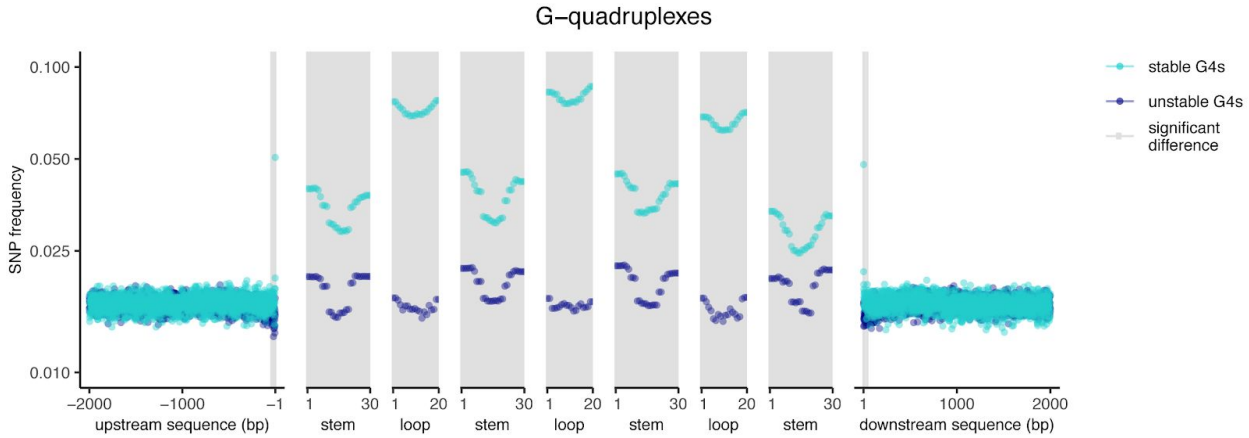
A



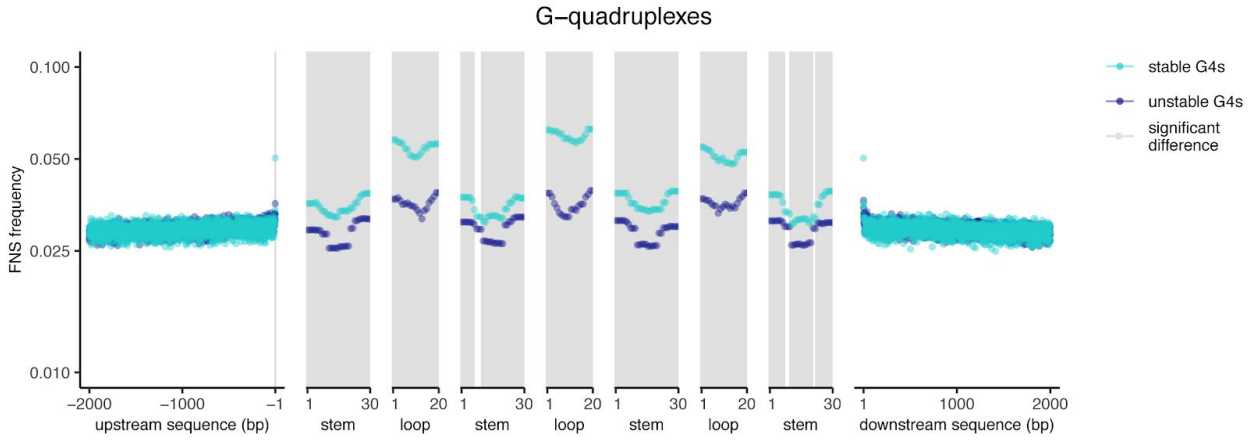
B



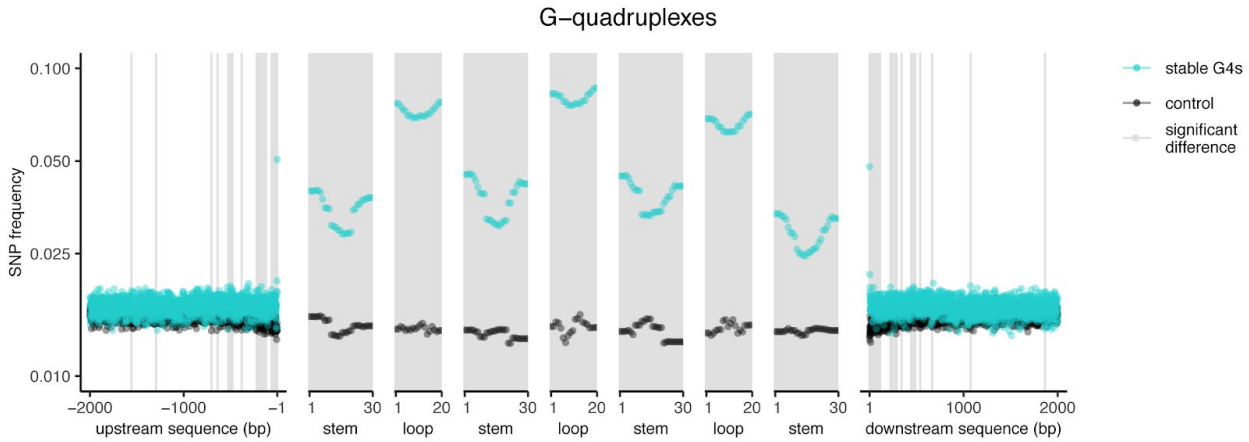
C


G−quadruplexes
SNP frequency
0.100
0.050
0.025
0.010
−2000
−1000
−1
upstream sequence (bp)
1
30
stem
1
20
loop
stem
loop
stem
loop
stem
1
1000
2000
downstream sequence (bp)
stable G4s
unstable G4s
significant difference

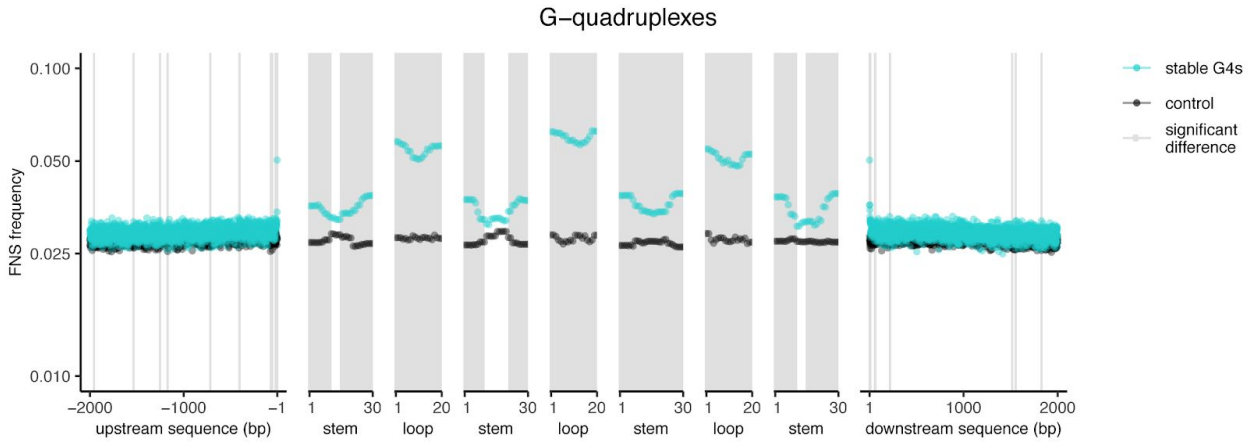
D


G−quadruplexes
FNS frequency
0.100
0.050
0.025
0.010
−2000
−1000
−1
upstream sequence (bp)
1
30
stem
1
20
loop
stem
loop
stem
loop
stem
1
1000
2000
downstream sequence (bp)
stable G4s
unstable G4s
significant difference

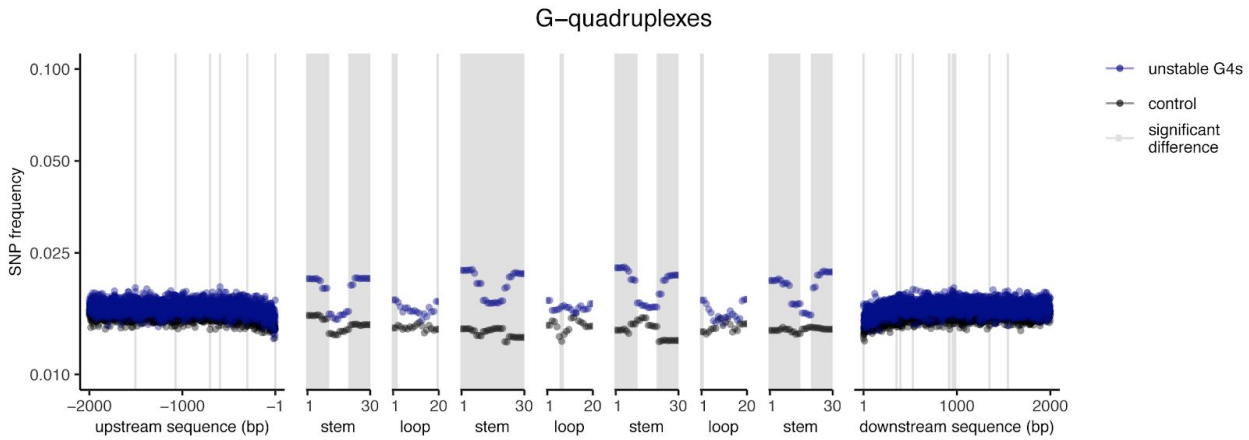
E


G−quadruplexes
SNP frequency
0.100
0.050
0.025
0.010
−2000
−1000
−1
upstream sequence (bp)
1
30
stem
1
20
loop
stem
loop
stem
loop
stem
1
1000
2000
downstream sequence (bp)
stable G4s
control
significant difference

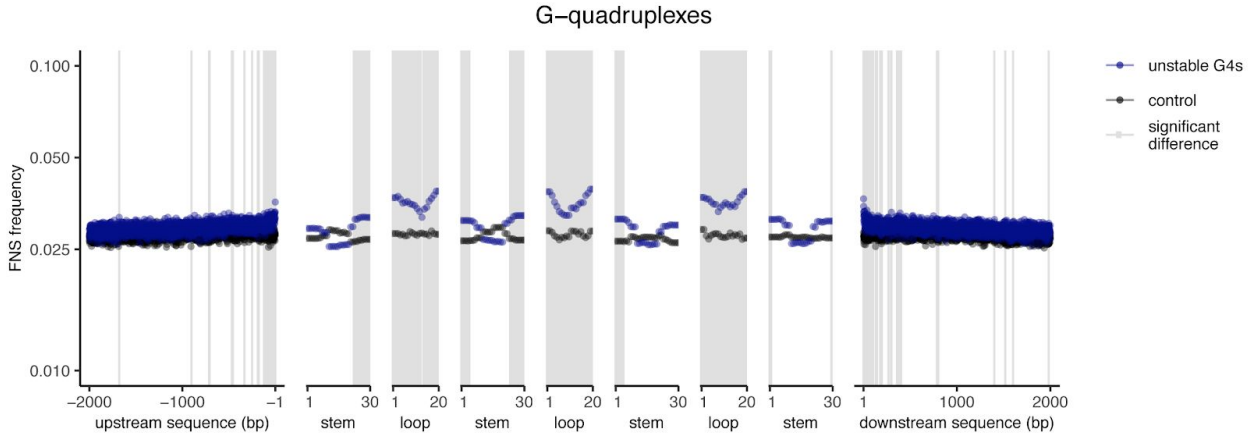
F


G–quadruplexes
FNS frequency
0.100
0.050
0.025
0.010
−2000
−1000
−1
upstream sequence (bp)
1
30
stem
1
20
loop
stem
loop
stem
loop
stem
1
1000
2000
downstream sequence (bp)
stable G4s
control
significant difference

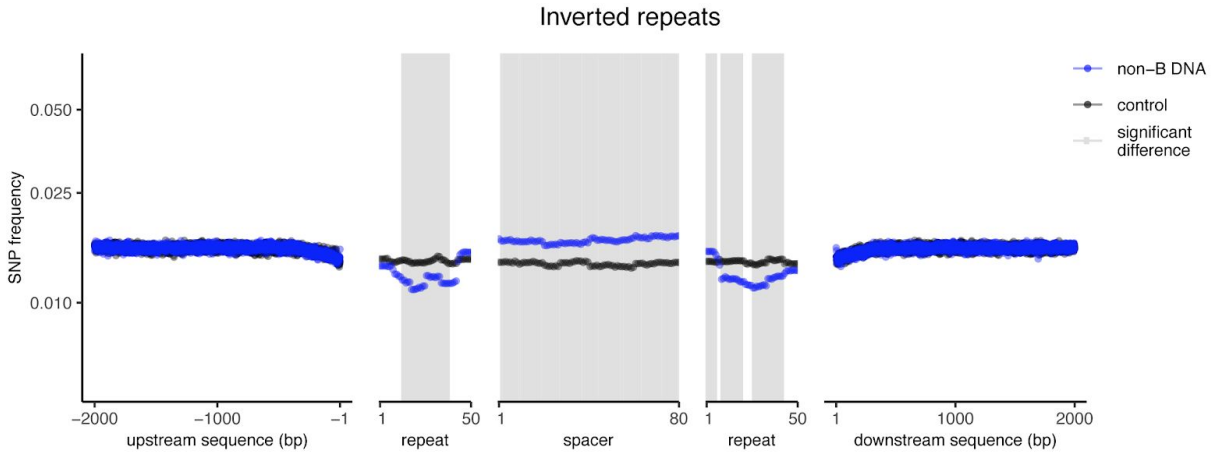
G


G–quadruplexes
SNP frequency
0.100
0.050
0.025
0.010
−2000
−1000
−1
upstream sequence (bp)
stem
loop
stem
loop
stem
loop
stem
downstream sequence (bp)
unstable G4s
control
significant difference

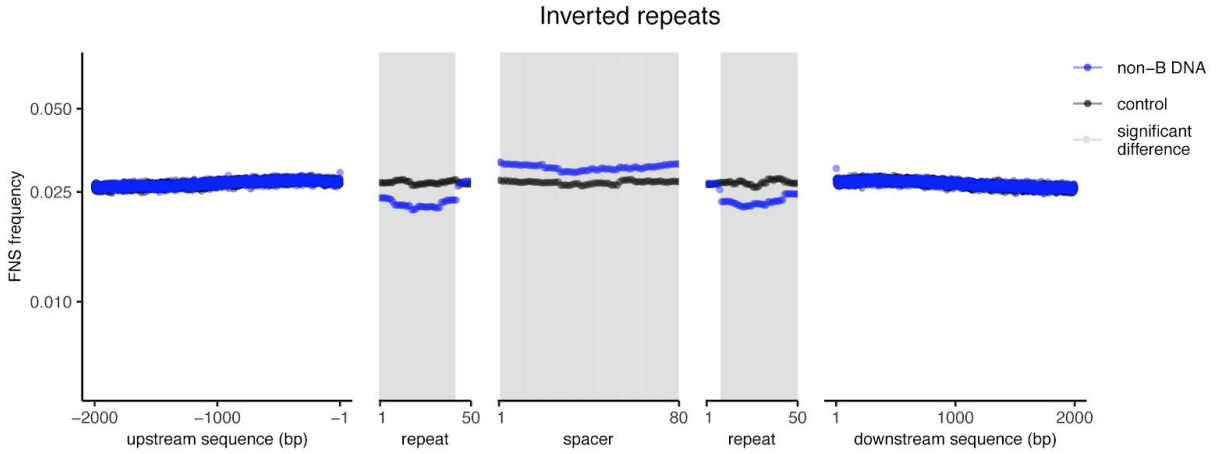
H


G–quadruplexes
FNS frequency
0.100
0.050
0.025
0.010
−2000
−1000
−1
upstream sequence (bp)
stem
loop
stem
loop
stem
loop
stem
downstream sequence (bp)
unstable G4s
control
significant difference

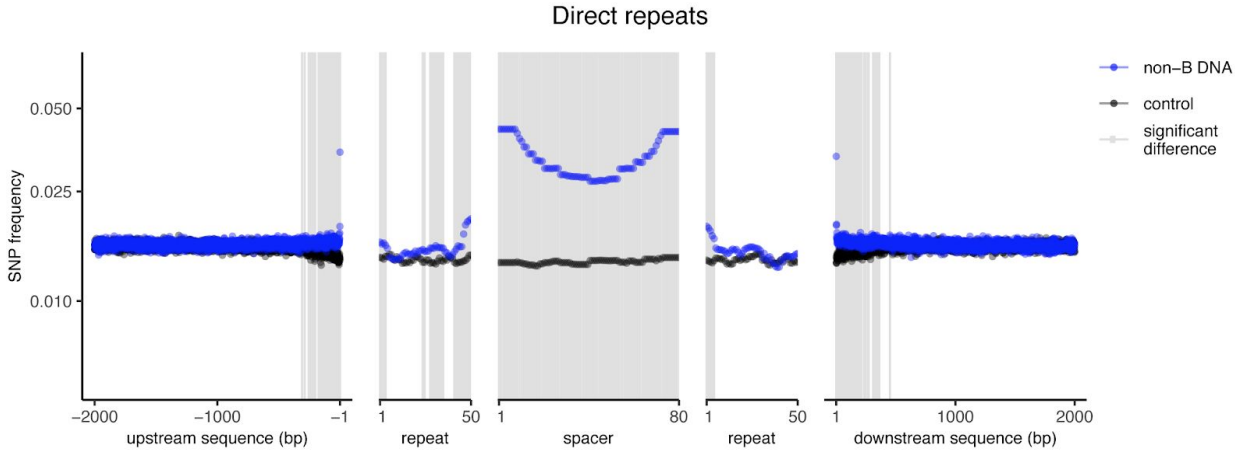
I


Inverted repeats
SNP frequency
0.050
0.025
0.010
−2000
−1000
−1
upstream sequence (bp)
1
50
repeat
1
80
spacer
1
50
repeat
1
1000
2000
downstream sequence (bp)
non−B DNA
control
significant difference

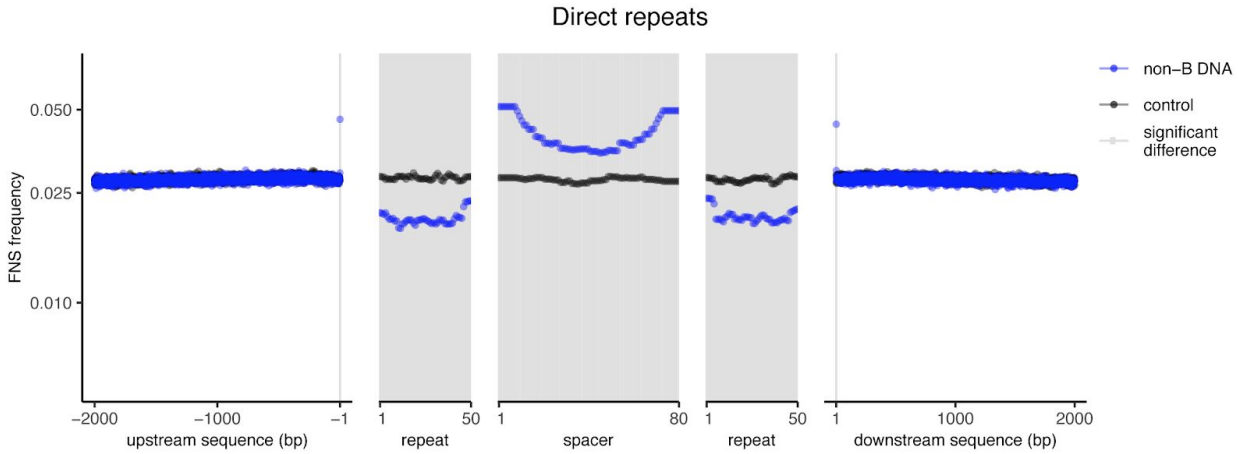
J


Inverted repeats
FNS frequency
0.050
0.025
0.010
−2000
−1000
−1
upstream sequence (bp)
1
50
repeat
1
80
spacer
1
50
repeat
1
1000
2000
downstream sequence (bp)
non−B DNA
control
significant difference

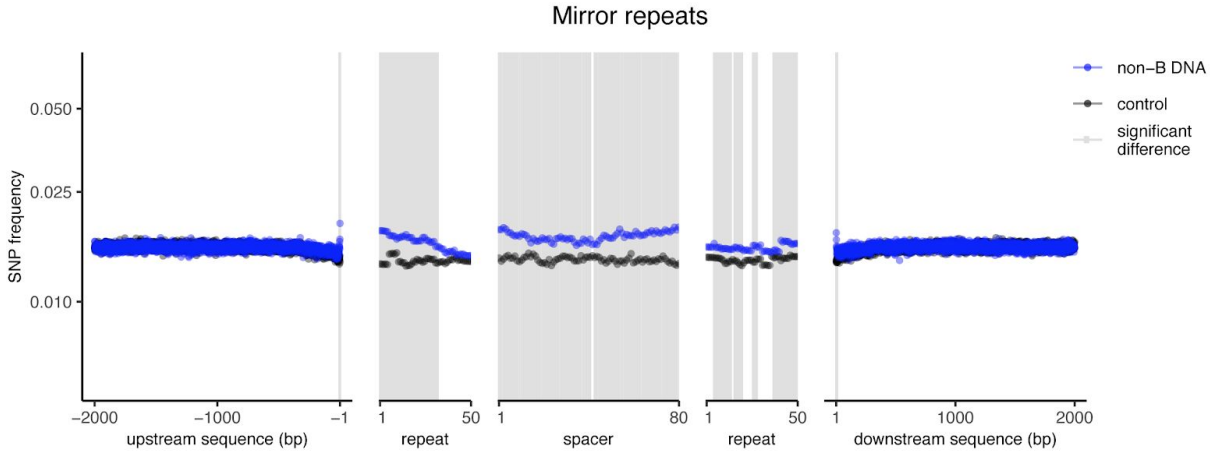
K


Direct repeats
SNP frequency
0.050
0.025
0.010
−2000
−1000
−1
upstream sequence (bp)
1
50
repeat
1
80
spacer
1
50
repeat
1
1000
2000
downstream sequence (bp)
non−B DNA
control
significant difference

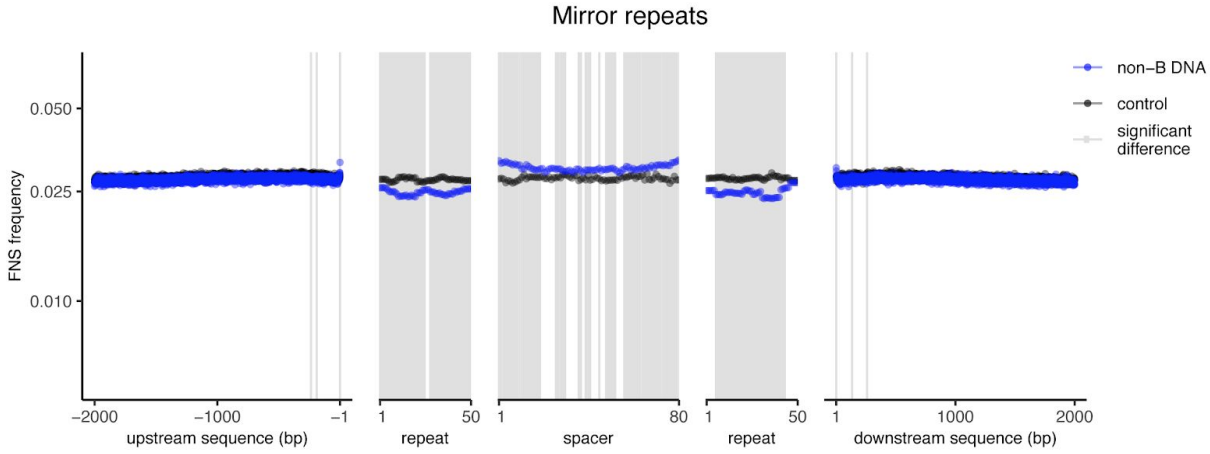
L


Direct repeats
FNS frequency
0.050
0.025
0.010
−2000
−1000
−1
upstream sequence (bp)
1
50
repeat
1
80
spacer
1
50
repeat
1
1000
2000
downstream sequence (bp)
non−B DNA
control
significant difference

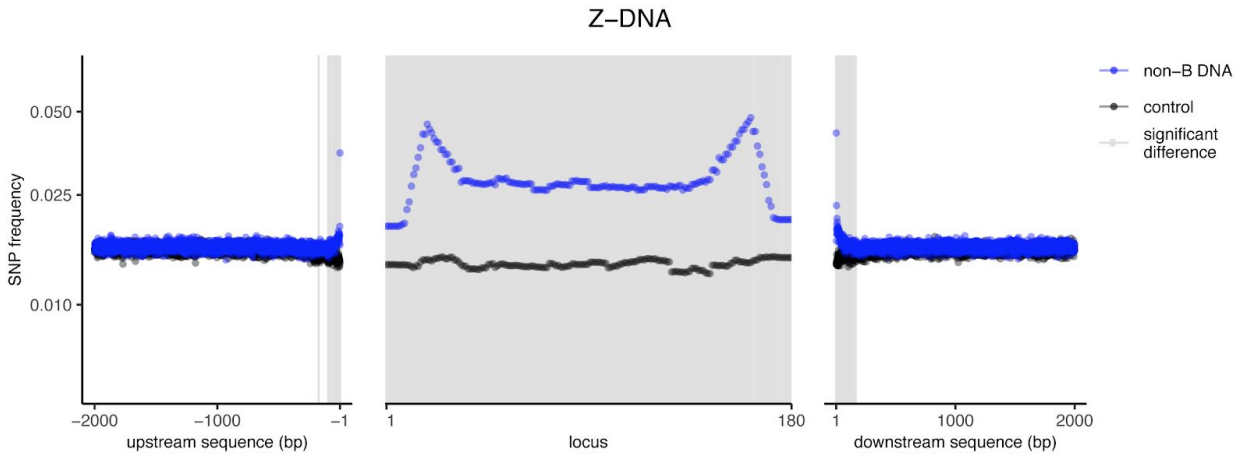
M


Mirror repeats
SNP frequency
0.050
0.025
0.010
−2000
−1000
−1
upstream sequence (bp)
1
50
repeat
1
80
spacer
1
50
repeat
1
1000
2000
downstream sequence (bp)
non−B DNA
control
significant difference

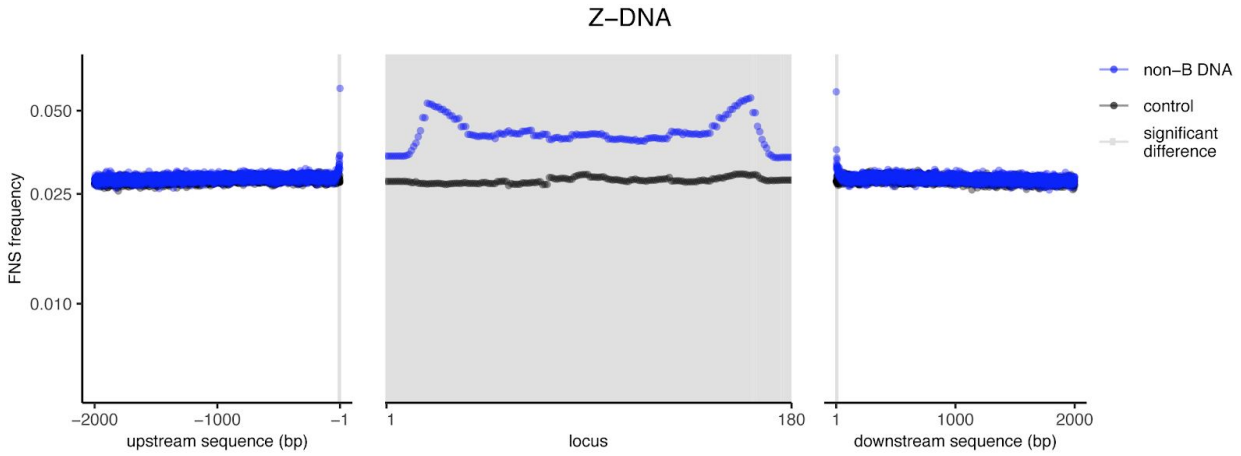
N


Mirror repeats
FNS frequency
0.050
0.025
0.010
−2000
−1000
−1
upstream sequence (bp)
1
50
repeat
1
80
spacer
1
50
repeat
1
1000
2000
downstream sequence (bp)
non−B DNA
control
significant difference

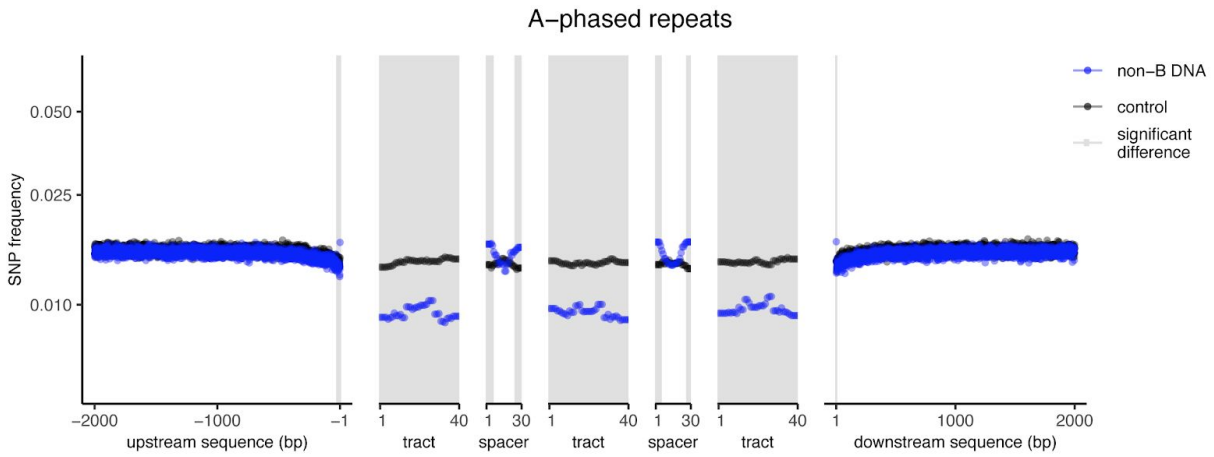
O


Z–DNA
SNP frequency
0.050
0.025
0.010
−2000
−1000
−1
upstream sequence (bp)
1
180
locus
1
1000
2000
downstream sequence (bp)
non–B DNA
control
significant difference

P


Z–DNA
FNS frequency
0.050
0.025
0.010
−2000
−1000
−1
upstream sequence (bp)
1
180
locus
1
1000
2000
downstream sequence (bp)
non–B DNA
control
significant difference

Q


A−phased repeats
SNP frequency
0.050
0.025
0.010
non−B DNA
control
significant difference
−2000
−1000
−1
1
40
1
30
1
40
1
30
1
40
1
1000
2000
upstream sequence (bp)
tract
spacer
tract
spacer
tract
downstream sequence (bp)

R

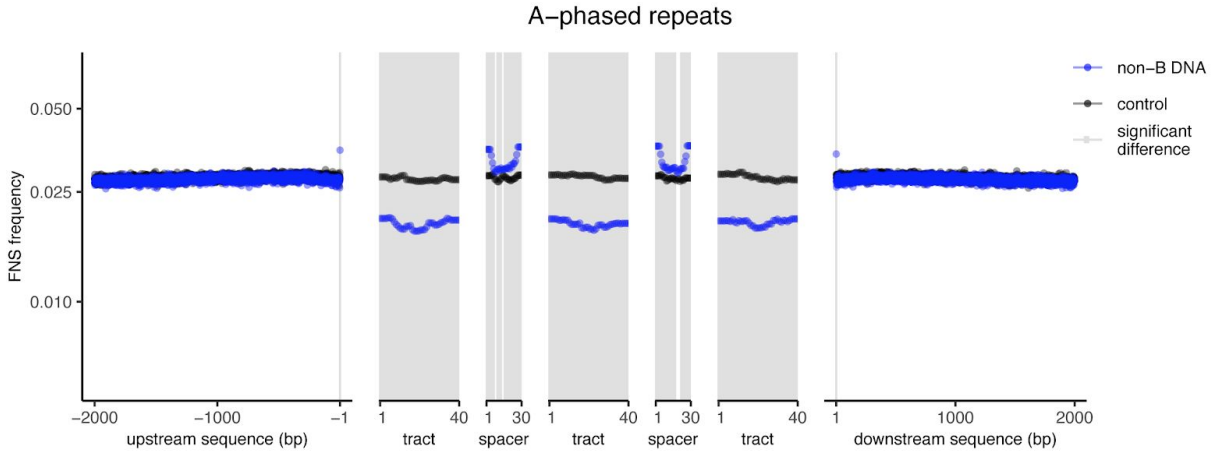
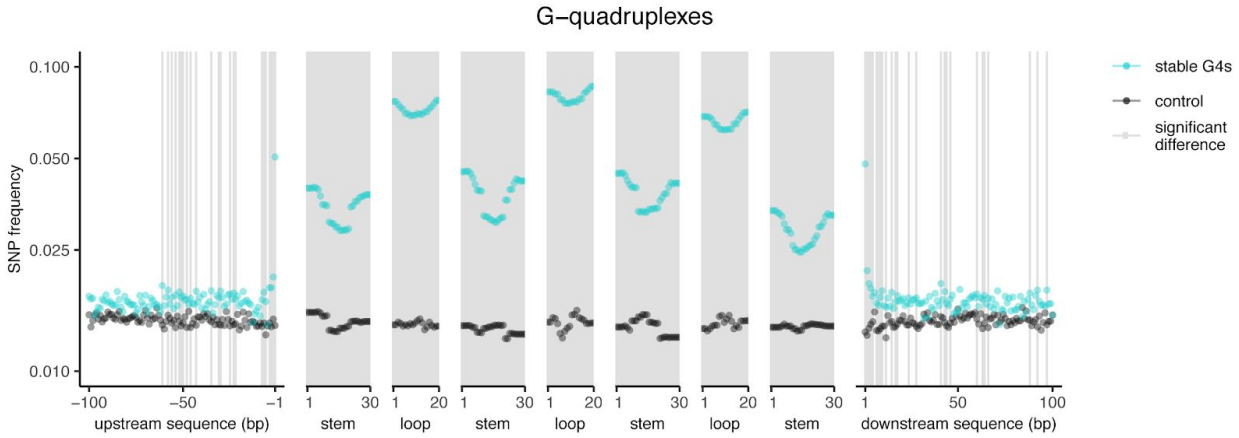
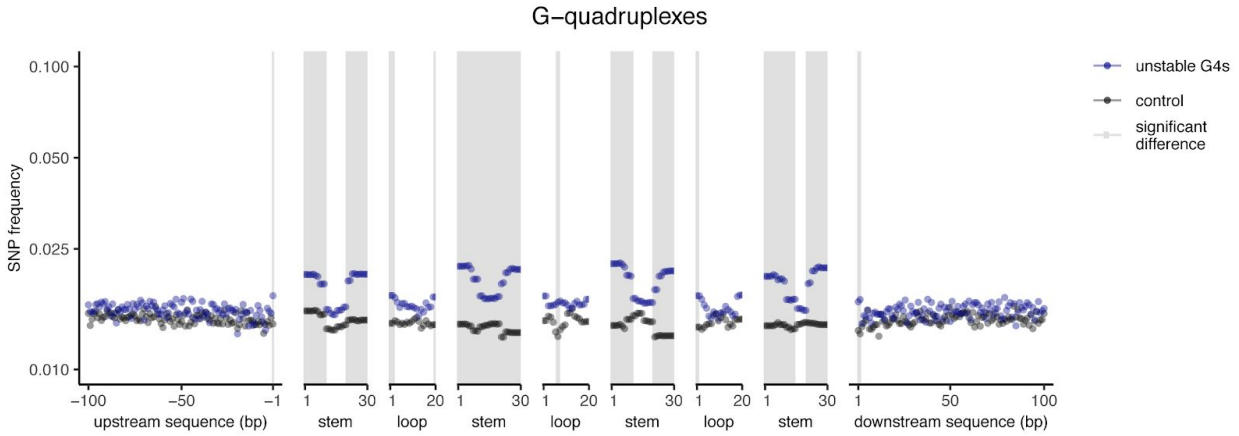

A−phased repeats
FNS frequency
0.050
0.025
0.010
non−B DNA
control
significant difference
−2000
−1000
−1
1
40
1
30
1
40
1
30
1
40
1
1000
2000
upstream sequence (bp)
tract
spacer
tract
spacer
tract
downstream sequence (bp)

Figure S9. Genome-wide nucleotide substitution frequencies at G4 loci. The positions of nucleotide substitutions within motifs were scaled based on motif size (see Materials and Methods for details). Stems are runs of guanines and loops are unspecified nucleotides between stems. Flanking regions are the 2 kilobases up- and downstream from the loci. For clarity of visualization, only the first 100 bps are shown (the full 2 kbs are shown in Fig. S2) and the Y-axes are displayed on a log scale. Gray areas indicate significantly different rates between groups (IWTomics adjusted *p*-value curve <0.01). A comparison for single-nucleotide polymorphism (SNP) frequencies between stable (**A**) and unstable (**B**) G4 loci. A comparison for fixed nucleotide substitution (FNS) frequencies between stable (**C**) and unstable (**D**) G4 loci.

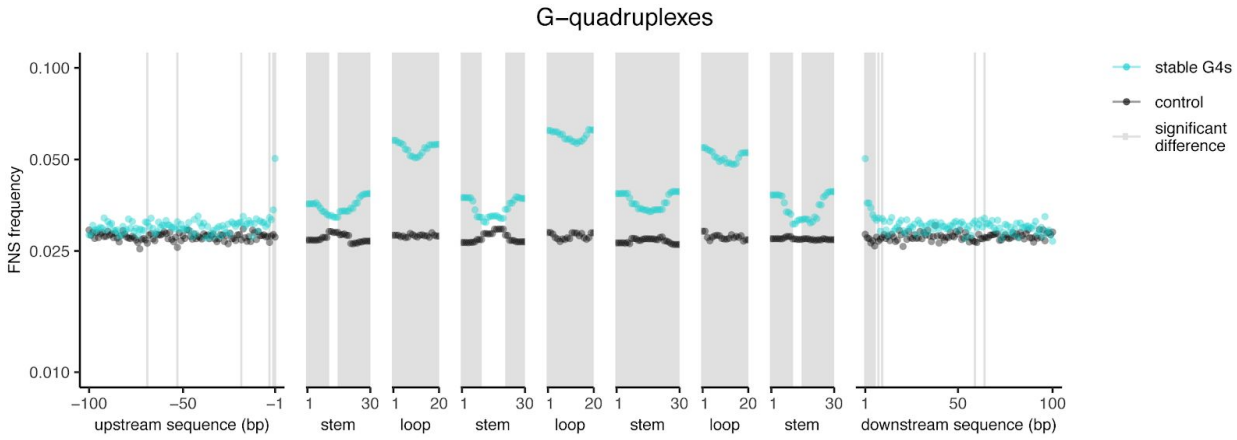
A Stable G4s vs. controls



B Unstable G4s vs. controls



C Stable G4s vs. controls



D Unstable G4s vs. controls

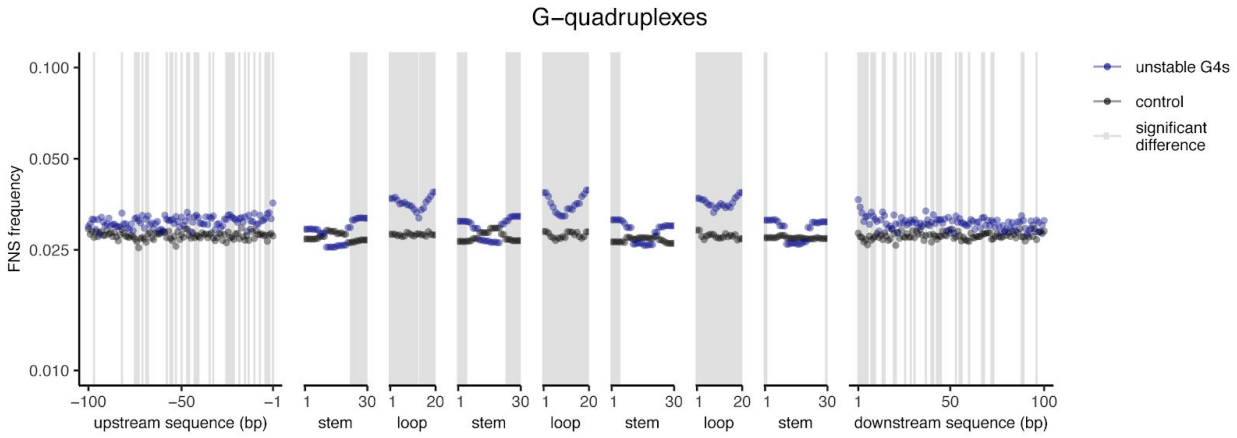


Figure S10. Frequencies of polymorphic substitutions at the immediate first 5' and 3' flanking positions of stable G4 loci annotated on the non-reference strand. The substitutions are shown with respect to the non-reference strand. Only the frequencies of trinucleotides present at the immediate flanking positions of stable G4 loci were compared with those present at control sequences (trinucleotides present only in control sequences were not considered). A correction for the trinucleotide context was applied (see Materials and Methods). Two-sided Fisher's exact test was used to evaluate significant differences, and *p*-values were adjusted for multiple testing using Bonferroni correction. An asterisk (*) marks significant differences between G4 and control sequences (adjusted *p*-value < 0.05).

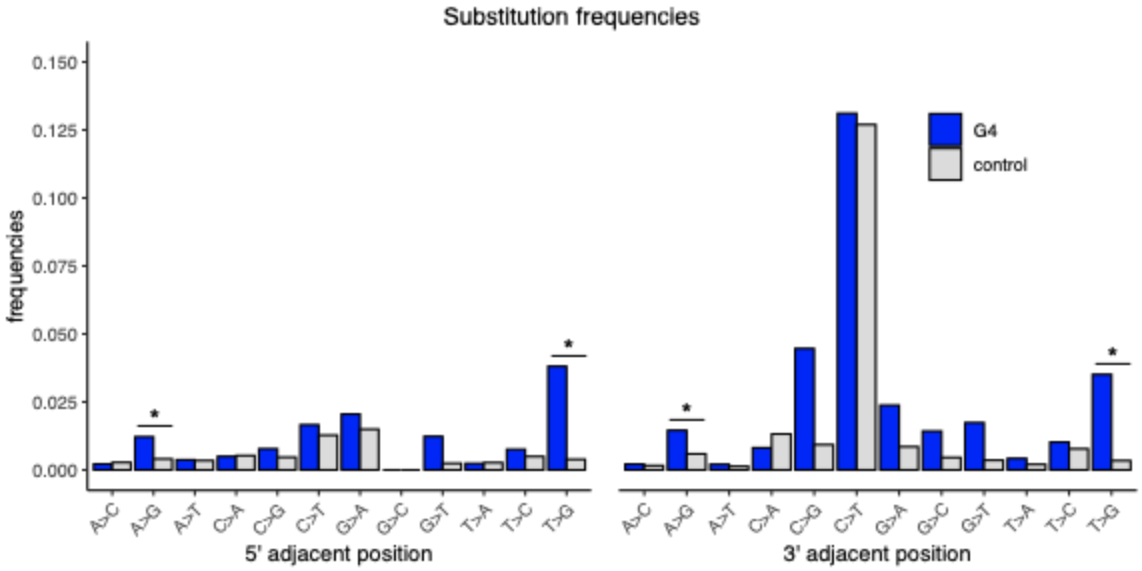


Figure S11. Relationships between SNP frequency and non-B DNA. Red curves represent loess fits (locally estimated scatterplot smoothing) shown to facilitate the visualization of trends.

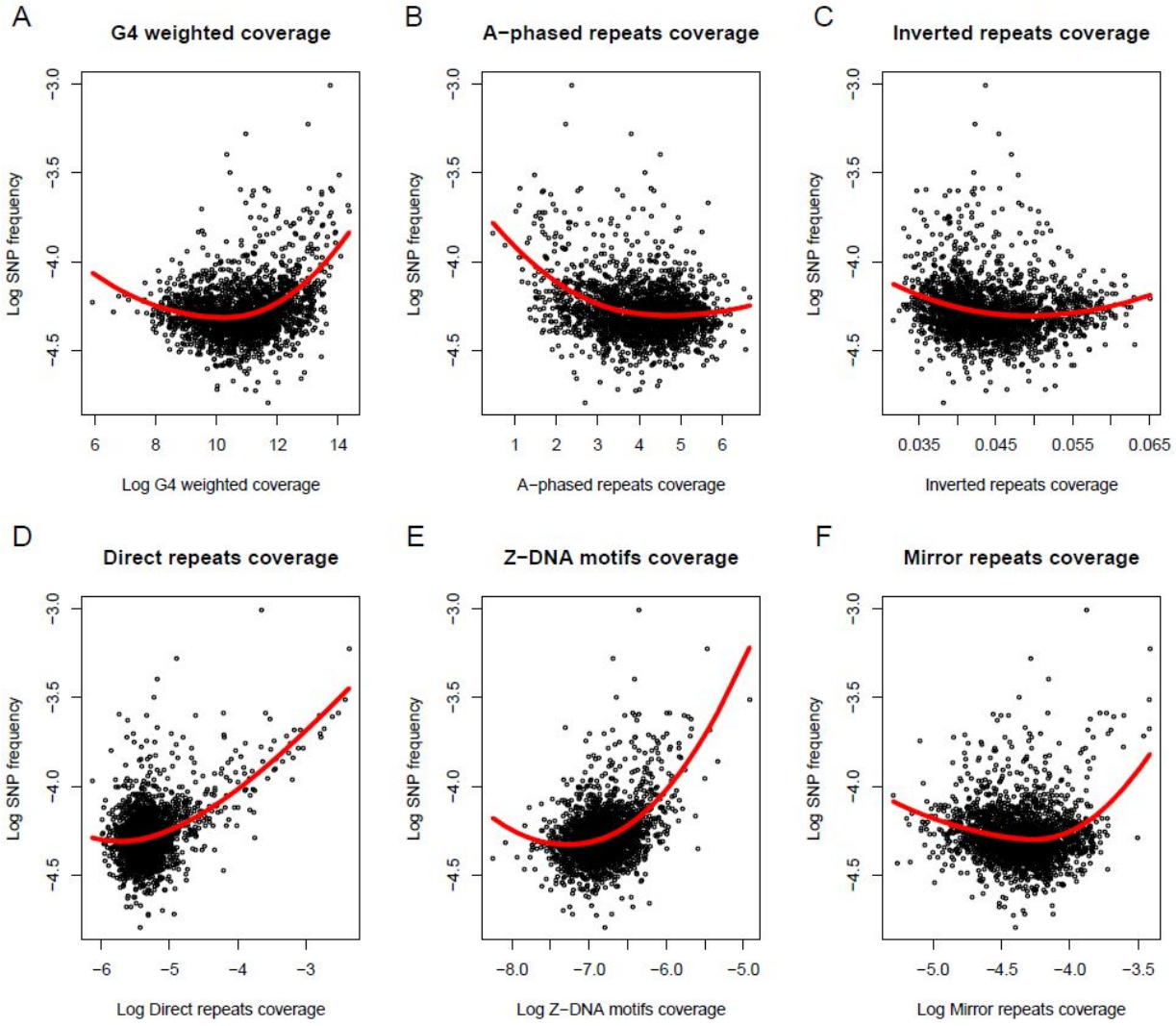


Figure S12. Site frequency spectra at stems and loops of G-quadruplexes analyzed in the NCNR subgenome. Two-sample Kolmogorov-Smirnov two-sided test *p*-value=0.