Non-canonical DNA in human and other ape telomere-to-telomere genomes

Linnéa Smeds, Kaivan Kamali, Kateryna D. Makova

Department of Biology, Penn State University, University Park, PA 16803

Correspondence to Kateryna Makova (kdm16@psu.edu)

Abstract

Non-canonical (non-B) DNA structures–bent DNA, slipped-strand structures, palindromes and hairpins, triplex DNA, G-quadruplexes, etc.–which form at certain sequence motifs (A-phased repeats, direct repeats, inverted repeats, mirror repeats, etc.), have recently emerged as important regulators of many cellular processes and as drivers of genome evolution. Yet, they have been understudied due to their highly repetitive nature and potentially inaccurate sequences when examined with short-read technologies. The recent availability of human and other ape telomere-to-telomere (T2T) genomes, which were generated with long reads, provides an opportunity to study the repertoire of non-canonical DNA motifs comprehensively. Here we characterized non-B DNA motifs in the T2T genomes of human and six non-human apes–bonobo, chimpanzee, gorilla, Bornean orangutan, Sumatran orangutan, and siamang. We found that non-B DNA motifs are enriched at the genomic regions that were added to these T2T assemblies as compared to previous assembly versions. Non-B DNA motifs occupied 9-15%, 9-11%, and 12-38% of autosomal, chromosome X, and chromosome Y sequences, respectively. Gorilla had the highest percentage of its genome occupied by non-B DNA motifs. The same DNA sequences were frequently annotated with multiple non-B DNA motif types. Several types of non-B DNA motifs had high densities at short arms of acrocentric chromosomes and non-canonical structures might contribute to satellite dynamics in these regions. Most centromeres showed an enrichment in at least one non-B DNA motif type, consistent with the role of non-B structures in determining centromeres. Our results highlight the uneven distribution of predicted non-B DNA structures across ape genomes and suggest their novel functions in the previously inaccessible genomic regions.

Introduction

In addition to canonical B DNA—the right-handed double helix with 10 base pairs per turn [1]—13% of the human genome at certain motifs was estimated to be able to fold into non-canonical (non-B) DNA structures [2]. Such non-B DNA conformations include cruciforms and hairpins formed by inverted repeats, triple helices (also called H-DNA) formed by mirror repeats, slipped strands formed by direct repeats, left-handed Z-DNA with 12 base pairs per turn formed by alternating purines and pyrimidines, and G-quadruplexes (G4s) formed by ≥4 'stems' consisting of ≥3 guanines and 'loops' consisting of any 1-7 bases (Fig. 1). Non-B DNA is universally present in species across the tree of life [2–7] forms in living cells [8,9], has been implicated in numerous cellular processes, and is frequently associated with diseases [10,11].

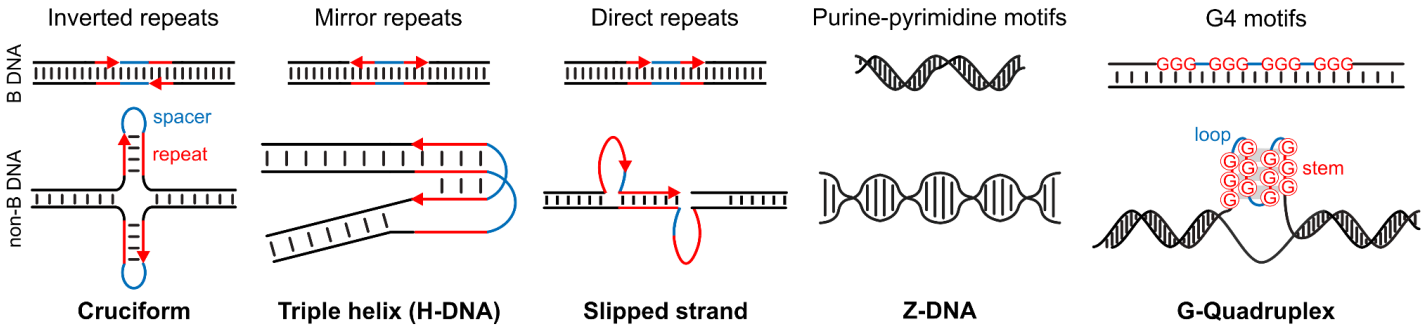

Figure 1. Types of non-B DNA. For structures formed from repeats: repeats are red, spacers are blue. For G4s: stems are red, loops are blue.

Non-B DNA sequence motifs range from tens to hundreds of nucleotides in length, and are present in tens of thousands of copies in the human genome [2]. Non-B DNA structure formation depends on cellular conditions: at any given time, DNA at a non-B motif can fold into either B or non-B form. For instance, folding into non-B forms is sensitive to temporal signals during cell differentiation and development [12,13], and is favored under oxidative stress [14,15] and after the induction of torsional stress behind DNA polymerases [16].

In part due to its transient nature, non-B DNA is increasingly recognized as a major regulator of myriad processes in the mammalian cell. Non-B DNA structures are involved in replication initiation [17,18]. G4s affect the life cycle of L1 transposable elements [19], and inverted repeats forming non-B DNA have been proposed to define centromeres [20]. G4 formation protects chromosome ends at telomeres [21]. Non-B DNA has been implicated in regulating transcription [3,22–33]. G4s are involved in regulating chromatin organization[33–38] and methylation of CpG islands [9]. The transcribed non-B DNA motifs can form structured RNA, which regulates alternative splicing [39], translation of mRNA [3,40,41], and function of non-coding RNA[42].

Notwithstanding their important functions, non-B structures may impede replication and elevate mutagenesis and genome instability. For instance, they can increase pausing and decrease the accuracy of replicative DNA polymerases *in vitro* [43–45]. To assist replication through such structures, the cell recruits error-prone specialized polymerases [46–49] and unwinding helicases [50]. Moreover, non-B DNA affects the efficiency of DNA repair pathways [9,44,51]. Increased mutagenesis and genomic instability at non-B DNA are evident in cancers with mutated components of these pathways [44]. In healthy cells, the effects of non-B DNA on replication progression, mutation rate, and genome instability remain controversial [52], nevertheless it has been suggested to be an important driver of genome evolution [53].

Non-B DNA structures have been implicated in neurodegenerative diseases (e.g. amyotrophic lateral sclerosis [54] and fragile X syndrome [55]). They are also the preferential sites of genome rearrangements [56] and affect gene expression in cancers [55]. Some diseases result from mutations in genes encoding proteins processing non-B DNA—e.g. Werner syndrome caused by the mutated WRN helicase unwinding G4s [57].

Despite their unequivocal importance for genome function, mutations, and diseases, non-B DNA has been challenging to study for several reasons. First, we and others have recently demonstrated that several sequencing technologies have increased error rates at non-B DNA motifs [58,59], with Illumina having higher error rates at such motifs than other technologies [58]. To overcome this limitation, the current recommendation is to

use multiple sequencing technologies as they differ in their biases at non-B DNA motifs [58]. Second, and relatedly, the incomplete genome assemblies have hindered the full characterization of non-B DNA motifs.

Here we characterized non-B DNA motif occurrences in the complete, telomere-to-telomere (T2T) genomes of human [60,61] and several non-human apes [62,63]. This approach has allowed us to provide a truly comprehensive view of non-B DNA distribution. Moreover, the human and primate T2T assemblies were produced with the combination of mostly long-read sequencing technologies thus minimizing the sequencing biases of each of them at non-B DNA motifs. Using this exhaustive dataset, we have addressed several questions that were impossible to answer prior to the availability of complete ape genomes, including the potential enrichment of non-B DNA at centromeres and repeats/satellites.

Results

Non-B DNA annotations. Most of non-B DNA motifs–A-phased repeats, direct repeats, inverted repeats, mirror repeats, short tandem repeats (STRs), and Z-DNA–were annotated in the latest versions of ape T2T genomes [63] with gfa (https://github.com/abcsFrederick/non-B_gfa). Some previous studies suggested that non-B DNA forms structures at inverted and mirror repeats only when the spacer length is below 15 bp [64–66], however this has been debated. We used the default parameters for the gfa annotations that allow for longer spacers (up to 100 bp) because of this uncertainty and because the majority of our annotations had spacers below 15 bp (Fig. S1). G4s were annotated with Quadron [67].

An overrepresentation of non-B DNA motifs in the newly added regions of the human T2T genomes. Notably, we observed an overrepresentation of most non-B DNA motif types in the newly added (or unalignable) sequences of the T2T human genome (CHM13) as compared to the previous, non-T2T version (hg38; Table 1), demonstrating the power of T2T genome assemblies in resolving these genomic regions. In particular, A-phased, direct, inverted, mirror repeats, and STRs were substantially overrepresented in the newly added sequences for the autosomes. Similar results were previously observed for great ape sex chromosomes [61,62] and autosomes [63].

Table 1. Non-B DNA motifs enriched at the newly added sequences of the human T2T genome, shown as fold enrichment in unaligned vs. aligned sequence as compared to hg38. Bold numbers indicate significantly different non-B DNA content in aligned vs. unaligned sequence (Chi-square goodness of fit with Bonferroni correction for multiple testing, *P*<0.001).

Chr/Non-B type	APR	DR	GQ	IR	MR	STR	Z-DNA	ALL
Autosomes	**6.72**	**18.75**	**0.96**	**1.63**	**2.61**	**13.44**	**0.65**	**4.98**
Chr X	**0.55**	**24.92**	**2.76**	**7.66**	**13.83**	**6.94**	0.98	**5.95**
Chr Y	**0.08**	**0.82**	**0.08**	**19.70**	**19.43**	**0.44**	**0.01**	**2.83**

Distribution of non-B DNA motifs between sex chromosomes and autosomes, and among species. The non-B DNA motif annotations in the T2T ape genomes revealed the complete picture of the distribution of these motifs among different chromosome types, non-B DNA types, and species (Table 2). We found that, depending on species, autosomes had 9.20-14.89% of their sequence annotated in non-B DNA motifs, whereas the X chromosome had a lower percentage (8.78-11.41%) and the Y chromosome had a substantially higher percentage (12.23-37.85%). As a rule, inverted, mirror, direct repeats, and STRs were more abundant than the other non-B DNA motif types. G-quadruplexes occupied 0.80-1.01% of autosomal sequences, and usually a lower percentage of sex chromosomal sequences. A-phased repeats and Z-DNA each occupied a lower percentage of autosomal sequences as compared to that for G-quadruplexes. Among species analyzed, gorilla had the highest percentage of its genome annotated in non-B DNA motifs, whereas siamang had the lowest. Some non-B DNA motif types were more abundant in some species but not the others. For instance, direct repeats were highly abundant in gorilla.

Table 2. Non-B DNA annotation in T2T genomes. Sum of non-B DNA motifs (in Mb and percentage of total length) for autosomes and sex chromosomes separately. APR: A-phased repeats, DR: direct repeats, GQ: G-quaruplexes, IR: inverted repeats, MR: mirror repeats, STR: short tandem repeats.

Chromosomes	Species	APR		DR		GQ		IR		MR		STR		Z-DNA		ALL	
Autosomes	Chimpanzee	11.8	0.40%	88.9	2.98%	26.7	0.89%	146.0	4.89%	87.3	2.92%	53.7	1.80%	6.6	0.22%	335.9	11.24%
	Bonobo	13.7	0.45%	109.3	3.60%	27.0	0.89%	155.4	5.12%	89.5	2.95%	70.9	2.34%	6.6	0.22%	367.7	12.11%
	Human	13.8	0.47%	81.5	2.81%	26.2	0.90%	138.2	4.76%	82.4	2.84%	65.1	2.24%	6.6	0.23%	320.2	11.04%
	Gorilla	11.8	0.36%	223.8	6.78%	33.3	1.01%	164.1	4.97%	103.2	3.13%	113.3	3.43%	6.7	0.20%	491.4	14.89%
	B. orangutan	12.2	0.40%	87.9	2.92%	26.7	0.89%	138.1	4.59%	82.6	2.74%	62.4	2.07%	6.5	0.22%	329.9	10.96%
	S. orangutan	12.0	0.39%	80.5	2.66%	25.7	0.85%	140.9	4.65%	84.3	2.78%	59.6	1.97%	6.6	0.22%	325.1	10.73%
	Siamang	10.4	0.34%	45.3	1.48%	24.4	0.80%	144.7	4.72%	73.2	2.39%	47.5	1.55%	6.0	0.20%	282.2	9.20%
Chr X	Chimpanzee	0.61	0.40%	2.51	1.63%	1.08	0.70%	7.19	4.67%	4.57	2.97%	2.37	1.54%	0.32	0.21%	14.43	9.37%

	Bonobo	0.62	0.39%	3.00	1.87%	1.08	0.68%	7.46	4.66%	4.86	3.03%	2.41	1.50%	0.34	0.21%	15.43	9.63%
	Human	0.61	0.40%	2.35	1.52%	1.03	0.67%	7.45	4.83%	4.52	2.93%	2.40	1.55%	0.34	0.22%	14.51	9.40%
	Gorilla	0.64	0.36%	7.75	4.36%	1.05	0.59%	7.53	4.24%	4.81	2.71%	2.44	1.37%	0.34	0.19%	20.26	11.41%
	B. orangutan	0.63	0.39%	2.35	1.46%	1.05	0.65%	7.60	4.72%	5.00	3.10%	2.44	1.51%	0.32	0.20%	15.24	9.47%
	S. orangutan	0.63	0.39%	2.35	1.45%	1.04	0.64%	7.70	4.74%	5.05	3.11%	2.46	1.52%	0.34	0.21%	15.40	9.47%
	Siamang	0.58	0.35%	2.19	1.32%	0.97	0.59%	7.64	4.61%	4.19	2.53%	2.34	1.41%	0.30	0.18%	14.54	8.78%
Chr Y	Chimpanzee	0.53	1.45%	3.71	10.18%	0.24	0.66%	1.47	4.03%	1.30	3.57%	2.55	6.99%	0.08	0.22%	6.80	18.66%
	Bonobo	0.63	1.33%	6.81	14.50%	0.29	0.62%	1.90	4.05%	1.67	3.56%	4.70	10.01%	0.09	0.19%	10.84	23.09%
	Human	1.89	3.03%	8.01	12.83%	0.21	0.34%	9.54	15.27%	7.64	12.24%	2.90	4.65%	0.07	0.12%	20.24	32.41%
	Gorilla	0.57	0.85%	11.97	17.76%	0.36	0.53%	9.27	13.75%	8.39	12.45%	6.82	10.12%	0.07	0.11%	25.52	37.85%
	B. orangutan	0.58	1.16%	4.16	8.42%	0.26	0.53%	2.06	4.16%	1.54	3.12%	2.23	4.52%	0.08	0.16%	8.40	17.01%
	S. orangutan	1.14	1.68%	8.14	12.00%	0.30	0.44%	2.72	4.02%	1.87	2.75%	4.15	6.13%	0.08	0.12%	14.16	20.87%
	Siamang	0.27	0.90%	1.26	4.22%	0.29	0.96%	1.27	4.25%	0.85	2.85%	1.05	3.50%	0.05	0.17%	3.66	12.23%

Overlapping annotations of different non-B DNA motif types. Importantly, we found substantial overlap among non-B DNA annotations of different types (Fig. 2). For instance, ~70%, ~55%, and ~48% of Z-DNA sequence annotations on the autosomes (and even higher percentages on sex chromosomes) overlapped with the STR, mirror repeat, and direct repeat annotations, respectively. This suggests alternative structure formation afforded by the same genomic sequence.

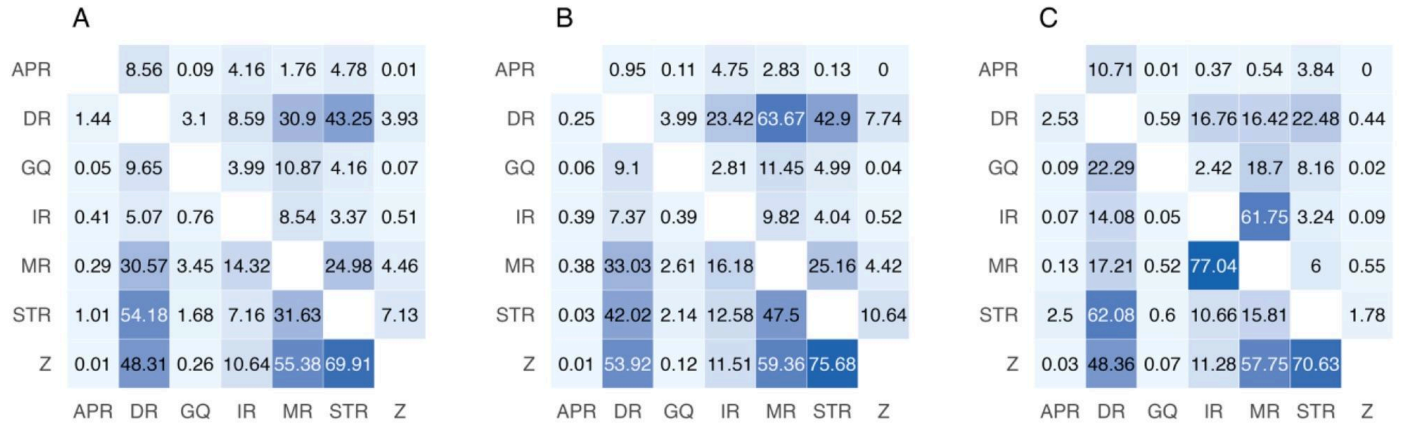

Figure 2. Overlap (i.e. the same bases annotated) between different non-B DNA motif types in human, on (A) autosomes, (B) chromosome X, and (C) chromosome Y, given as percentage of the left type that overlap with the bottom type. APR: A-phased repeats; DR: direct repeats; GQ: G-quadruplexes; IR: inverted repeats; MR: mirrored repeats; STR: short tandem repeats; Z: Z-DNA.

Distribution of non-B DNA motifs along the chromosomes. When visually inspecting the variation in the density of non-B DNA motifs along the chromosomes (Fig. 3), we discovered the following patterns. In human (Fig. 3C), for instance, we found high densities of non-B DNA motifs–G-quadruplexes, inverted repeats, mirror repeats, and STRs–on the short arms of acrocentric chromosomes (chromosomes 13, 14, 15, 21, and 22), and in the heterochromatic region of the Y chromosome. G-quadruplexes frequently had high densities at the subtelomeric regions. An interesting pattern was observed at the centromeres. At some chromosomes, non-B DNA had high density within the centromeres themselves, for instance, inverted repeats and mirror repeats for chromosome 3 and inverted repeats and A-phased repeats for chromosome 11. At some other chromosomes, non-B DNA motifs had high density immediately adjacent to the centromere, e.g., for chromosomes 1 and 9. Similar patterns were observed for other great apes (Fig. 3).

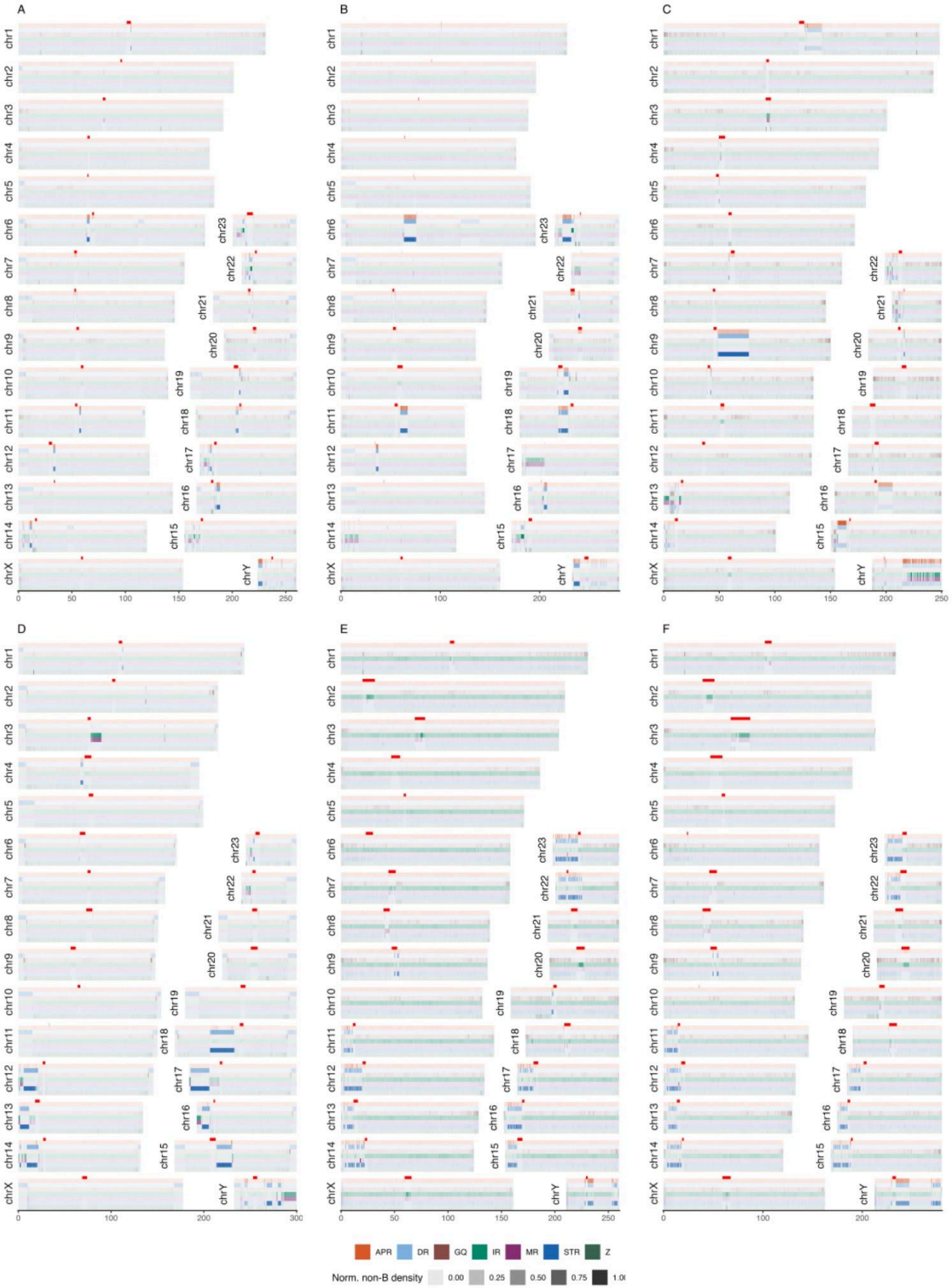

Figure 3. Non-B density along the chromosomes for (A) chimpanzee, (B) bonobo, (C) human, (D) gorilla, (E) Bornean orangutan, and (F) Sumatran orangutan, normalized to a scale from 0-1 for each non-B DNA type separately. Active centromeres are marked with red stripes

above the non-B DNA tracks. Note that an active centromere was not found for chr10 in Sumatran orangutan. Abbreviations for non-B DNA are as in Fig. 2.

Enrichment of non-B DNA motifs at centromeres. We performed a more detailed analysis (Fig. 4) of enrichment of non-B DNA at experimentally annotated, active centromeres of apes in our data set [63], as it was suggested that non-B DNA determines centromere formation [20]. G-quadruplexes were always underrepresented at the centromeres. The other types of non-B DNA displayed species- and chromosome-specific trends. Inverted repeats displayed a moderate enrichment in approximately half of all centromeres. A-phased repeats showed enrichment at centromeres of some chromosomes in most species, with a particularly high enrichment and a large number of chromosomes affected for chimpanzee and human. An enrichment in Z-DNA was observed for centromeres at some chromosomes in bonobo, chimpanzee, and human, whereas an enrichment in mirror repeats was particularly evident for centromeres at some chromosomes in the two orangutan species. Most of centromeres (107/148, or >70%) showed enrichment in at least one type of non-B DNA, and over a quarter of centromeres (42/148, or 28%) having >2-fold enrichment (Fig. 4).

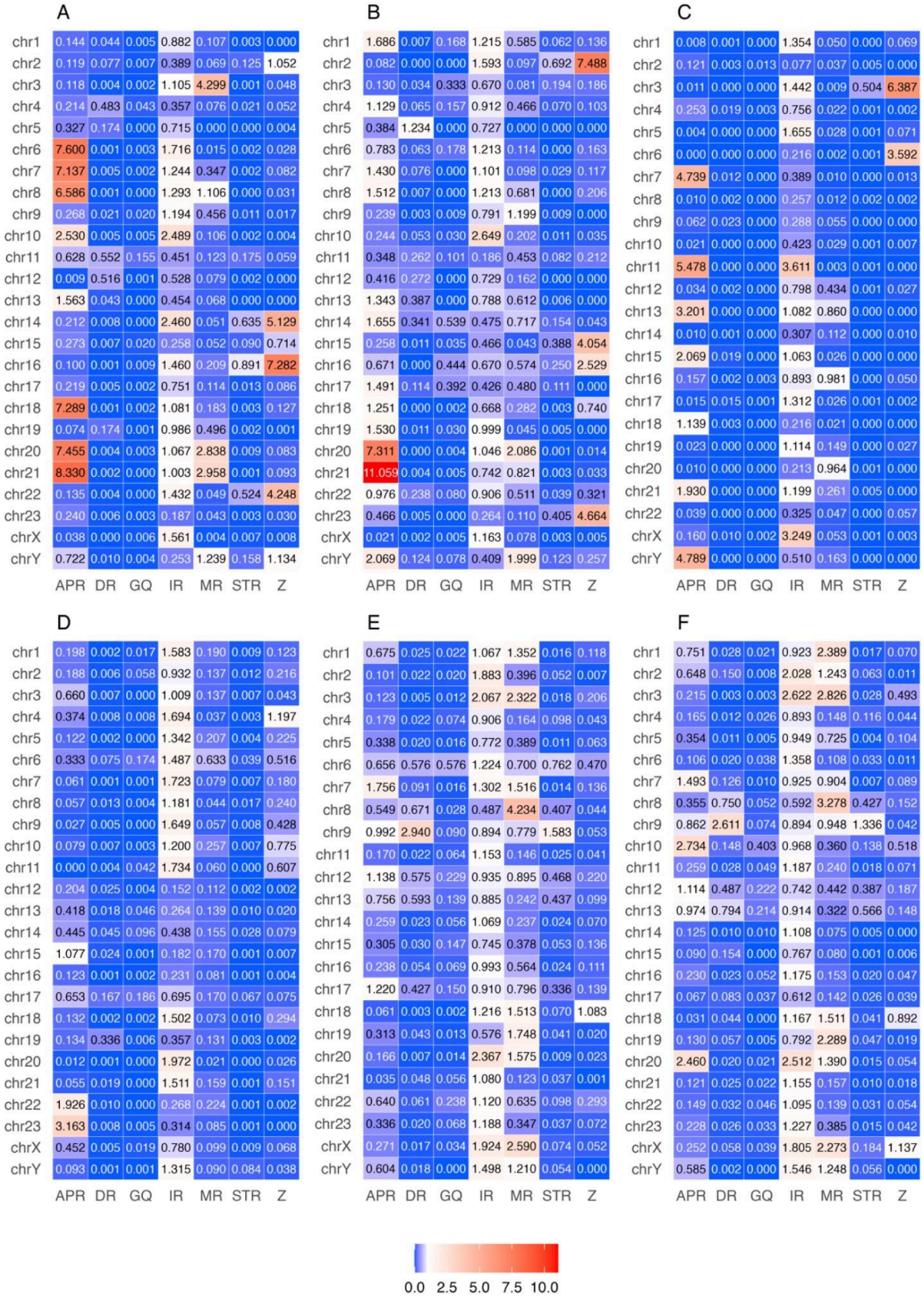
A
chr1 0.144 0.044 0.005 0.882 0.107 0.003 0.000
chr2 0.119 0.077 0.007 0.389 0.069 0.125 1.052
chr3 0.118 0.004 0.002 1.105 4.299 0.001 0.048
chr4 0.214 0.483 0.043 0.357 0.076 0.021 0.052
chr5 0.327 0.174 0.000 0.715 0.000 0.000 0.004
chr6 7.600 0.001 0.003 1.716 0.015 0.002 0.028
chr7 7.137 0.005 0.002 1.244 0.347 0.002 0.082
chr8 6.586 0.001 0.000 1.293 1.106 0.000 0.031
chr9 0.268 0.021 0.020 1.194 0.456 0.011 0.017
chr10 2.530 0.005 0.005 2.489 0.106 0.002 0.004
chr11 0.628 0.552 0.155 0.451 0.123 0.175 0.059
chr12 0.009 0.516 0.001 0.528 0.079 0.002 0.000
chr13 1.563 0.043 0.000 0.454 0.068 0.000 0.000
chr14 0.212 0.008 0.000 2.460 0.051 0.635 5.129
chr15 0.273 0.007 0.020 0.258 0.052 0.090 0.714
chr16 0.100 0.001 0.009 1.460 0.209 0.891 7.282
chr17 0.219 0.005 0.002 0.751 0.114 0.013 0.086
chr18 7.289 0.001 0.002 1.081 0.183 0.003 0.127
chr19 0.074 0.174 0.001 0.986 0.496 0.002 0.001
chr20 7.455 0.004 0.003 1.067 2.838 0.009 0.083
chr21 8.330 0.002 0.000 1.003 2.958 0.001 0.093
chr22 0.135 0.004 0.000 1.432 0.049 0.524 4.248
chr23 0.240 0.006 0.003 0.187 0.043 0.003 0.030
chrX 0.038 0.000 0.006 1.561 0.004 0.007 0.008
chrY 0.722 0.010 0.004 0.253 1.239 0.158 1.134
APR DR GQ IR MR STR Z
B
chr1 1.686 0.007 0.168 1.215 0.585 0.062 0.136
chr2 0.082 0.000 0.000 1.593 0.097 0.692 7.488
chr3 0.130 0.034 0.333 0.670 0.081 0.194 0.186
chr4 1.129 0.065 0.157 0.912 0.466 0.070 0.103
chr5 0.384 1.234 0.000 0.727 0.000 0.000 0.000
chr6 0.783 0.063 0.178 1.213 0.114 0.000 0.163
chr7 1.430 0.076 0.000 1.101 0.098 0.029 0.117
chr8 1.512 0.007 0.000 1.213 0.681 0.000 0.206
chr9 0.239 0.003 0.009 0.791 1.199 0.009 0.000
chr10 0.244 0.053 0.030 2.649 0.202 0.011 0.035
chr11 0.348 0.262 0.101 0.186 0.453 0.082 0.212
chr12 0.416 0.272 0.000 0.729 0.162 0.000 0.000
chr13 1.343 0.387 0.000 0.788 0.612 0.006 0.000
chr14 1.655 0.341 0.539 0.475 0.717 0.154 0.043
chr15 0.258 0.011 0.035 0.466 0.043 0.388 4.054
chr16 0.671 0.000 0.444 0.670 0.574 0.250 2.529
chr17 1.491 0.114 0.392 0.426 0.480 0.111 0.000
chr18 1.251 0.000 0.002 0.668 0.282 0.003 0.740
chr19 1.530 0.011 0.030 0.999 0.045 0.005 0.000
chr20 7.311 0.000 0.004 1.046 2.086 0.001 0.014
chr21 11.059 0.004 0.005 0.742 0.821 0.003 0.033
chr22 0.976 0.238 0.080 0.906 0.511 0.039 0.321
chr23 0.466 0.005 0.000 0.264 0.110 0.405 4.664
chrX 0.021 0.002 0.005 1.163 0.078 0.003 0.005
chrY 2.069 0.124 0.078 0.409 1.999 0.123 0.257
APR DR GQ IR MR STR Z
C
chr1 0.008 0.001 0.000 1.354 0.050 0.000 0.069
chr2 0.121 0.003 0.013 0.077 0.037 0.005 0.000
chr3 0.011 0.000 0.000 1.442 0.009 0.504 6.387
chr4 0.253 0.019 0.003 0.756 0.022 0.001 0.002
chr5 0.004 0.000 0.000 1.655 0.028 0.001 0.071
chr6 0.000 0.000 0.000 0.216 0.002 0.001 3.592
chr7 4.739 0.012 0.000 0.389 0.010 0.000 0.013
chr8 0.010 0.002 0.000 0.257 0.012 0.002 0.002
chr9 0.062 0.023 0.000 0.288 0.055 0.000 0.000
chr10 0.021 0.000 0.000 0.423 0.029 0.001 0.007
chr11 5.478 0.000 0.000 3.611 0.003 0.001 0.000
chr12 0.034 0.002 0.000 0.798 0.434 0.001 0.027
chr13 3.201 0.000 0.000 1.082 0.860 0.000 0.000
chr14 0.010 0.001 0.000 0.307 0.112 0.000 0.010
chr15 2.069 0.019 0.000 1.063 0.026 0.000 0.000
chr16 0.157 0.002 0.003 0.893 0.981 0.000 0.050
chr17 0.015 0.015 0.001 1.312 0.026 0.001 0.002
chr18 1.139 0.003 0.000 0.216 0.021 0.000 0.000
chr19 0.023 0.000 0.000 1.114 0.149 0.000 0.027
chr20 0.010 0.000 0.000 0.213 0.964 0.001 0.000
chr21 1.930 0.000 0.000 1.199 0.261 0.005 0.000
chr22 0.039 0.000 0.000 0.325 0.047 0.000 0.057
chrX 0.160 0.010 0.002 3.249 0.053 0.001 0.003
chrY 4.789 0.000 0.000 0.510 0.163 0.000 0.000
APR DR GQ IR MR STR Z
D
chr1 0.198 0.002 0.017 1.583 0.190 0.009 0.123
chr2 0.188 0.006 0.058 0.932 0.137 0.012 0.216
chr3 0.660 0.007 0.000 1.009 0.137 0.007 0.043
chr4 0.374 0.008 0.008 1.694 0.037 0.003 1.197
chr5 0.122 0.002 0.000 1.342 0.207 0.004 0.225
chr6 0.333 0.075 0.174 1.487 0.633 0.039 0.516
chr7 0.061 0.001 0.001 1.723 0.079 0.007 0.180
chr8 0.057 0.013 0.004 1.181 0.044 0.017 0.240
chr9 0.027 0.005 0.000 1.649 0.057 0.008 0.428
chr10 0.079 0.007 0.003 1.200 0.257 0.007 0.775
chr11 0.000 0.004 0.042 1.734 0.060 0.000 0.607
chr12 0.204 0.025 0.004 0.152 0.112 0.002 0.002
chr13 0.418 0.018 0.046 0.264 0.139 0.010 0.020
chr14 0.445 0.045 0.096 0.438 0.155 0.028 0.079
chr15 1.077 0.024 0.001 0.182 0.170 0.001 0.007
chr16 0.123 0.001 0.002 0.231 0.081 0.001 0.004
chr17 0.653 0.167 0.186 0.695 0.170 0.067 0.075
chr18 0.132 0.002 0.002 1.502 0.073 0.010 0.294
chr19 0.134 0.336 0.006 0.357 0.131 0.003 0.002
chr20 0.012 0.001 0.000 1.972 0.021 0.000 0.026
chr21 0.055 0.019 0.000 1.511 0.159 0.001 0.151
chr22 1.926 0.010 0.000 0.268 0.224 0.001 0.002
chr23 3.163 0.008 0.005 0.314 0.085 0.001 0.000
chrX 0.452 0.005 0.019 0.780 0.099 0.009 0.068
chrY 0.093 0.001 0.001 1.315 0.090 0.084 0.038
APR DR GQ IR MR STR Z
E
chr1 0.675 0.025 0.022 1.067 1.352 0.016 0.118
chr2 0.101 0.022 0.020 1.883 0.396 0.052 0.007
chr3 0.123 0.005 0.012 2.067 2.322 0.018 0.206
chr4 0.179 0.022 0.074 0.906 0.164 0.098 0.043
chr5 0.338 0.020 0.016 0.772 0.389 0.011 0.063
chr6 0.656 0.576 0.576 1.224 0.700 0.762 0.470
chr7 1.756 0.091 0.016 1.302 1.516 0.014 0.136
chr8 0.549 0.671 0.028 0.487 4.234 0.407 0.044
chr9 0.992 2.940 0.090 0.894 0.779 1.583 0.053
chr11 0.170 0.022 0.064 1.153 0.146 0.025 0.041
chr12 1.138 0.575 0.229 0.935 0.895 0.468 0.220
chr13 0.756 0.593 0.139 0.885 0.242 0.437 0.099
chr14 0.259 0.023 0.056 1.069 0.237 0.024 0.070
chr15 0.305 0.030 0.147 0.745 0.378 0.053 0.136
chr16 0.238 0.054 0.069 0.993 0.564 0.024 0.111
chr17 1.220 0.427 0.150 0.910 0.796 0.336 0.139
chr18 0.061 0.003 0.002 1.216 1.513 0.070 1.083
chr19 0.313 0.043 0.013 0.576 1.748 0.041 0.020
chr20 0.166 0.007 0.014 2.367 1.575 0.009 0.023
chr21 0.035 0.048 0.056 1.080 0.123 0.037 0.001
chr22 0.640 0.061 0.238 1.120 0.635 0.098 0.293
chr23 0.336 0.020 0.068 1.188 0.347 0.037 0.072
chrX 0.271 0.017 0.034 1.924 2.590 0.074 0.052
chrY 0.604 0.018 0.000 1.498 1.210 0.054 0.000
APR DR GQ IR MR STR Z
F
chr1 0.751 0.028 0.021 0.923 2.389 0.017 0.070
chr2 0.648 0.150 0.008 2.028 1.243 0.063 0.011
chr3 0.215 0.003 0.003 2.622 2.826 0.028 0.493
chr4 0.165 0.012 0.026 0.893 0.148 0.116 0.044
chr5 0.354 0.011 0.005 0.949 0.725 0.004 0.104
chr6 0.106 0.020 0.038 1.358 0.108 0.033 0.011
chr7 1.493 0.126 0.010 0.925 0.904 0.007 0.089
chr8 0.355 0.750 0.052 0.592 3.278 0.427 0.152
chr9 0.862 2.611 0.074 0.894 0.948 1.336 0.042
chr10 2.734 0.148 0.403 0.968 0.360 0.138 0.518
chr11 0.259 0.028 0.049 1.187 0.240 0.018 0.071
chr12 1.114 0.487 0.222 0.742 0.442 0.387 0.187
chr13 0.974 0.794 0.214 0.914 0.322 0.566 0.148
chr14 0.125 0.010 0.010 1.108 0.075 0.005 0.000
chr15 0.090 0.154 0.000 0.767 0.080 0.001 0.006
chr16 0.230 0.023 0.052 1.175 0.153 0.020 0.047
chr17 0.067 0.083 0.037 0.612 0.142 0.026 0.039
chr18 0.031 0.044 0.000 1.167 1.511 0.041 0.892
chr19 0.130 0.057 0.005 0.792 2.289 0.047 0.019
chr20 2.460 0.020 0.021 2.512 1.390 0.015 0.054
chr21 0.121 0.025 0.022 1.155 0.157 0.010 0.018
chr22 0.149 0.032 0.046 1.095 0.139 0.031 0.054
chr23 0.228 0.026 0.033 1.227 0.385 0.015 0.042
chrX 0.252 0.058 0.039 1.805 2.273 0.184 1.137
chrY 0.585 0.002 0.000 1.546 1.248 0.056 0.000
APR DR GQ IR MR STR Z
0.0 2.5 5.0 7.5 10.0

Figure 4. Fold-enrichment of non-B DNA densities in centromeres as compared to genome-wide densities for A) chimpanzee, B) bonobo, C) human, D) gorilla, E) Bornean orangutan and F) Sumatran orangutan. Underrepresentation of non-B DNA (values below 1) are shown in blue, while enrichment (values above 1) are shown in red. Abbreviations for non-B DNA are as in Fig. 2.

Enrichment of non-B DNA motifs upstream of transcription start sites (TSSs). Many types of non-B DNA were implicated in the regulation of transcription (see references in the Introduction). Consistent with these studies, but now analyzing the complete, T2T human genome, we found 5.53-, 1.81-, and 1.22-fold enrichment of G-quadruplexes, Z-DNA, and mirror repeats at putative promoter regions of human genes, defined as 1 kb upstream of TSSs (Table 3). The fold enrichment for the whole genome was significantly different from randomly selected genomic regions for all non-B DNA types except inverted repeats (two-sided permutation test with 10,000 permutations, $p<0.01$, after correcting for multiple testing).

Table 3. Enrichment of non-B DNA in human 1 kb upstream transcription start sites (TSSs), given as the fold change in density compared to genome-wide.

Region	APR	DR	GQ	IR	MR	STR	Z
autosomes	0.44	0.60	5.51	1.04	1.22	0.81	1.79
chrX	0.54	0.71	5.56	0.98	1.28	0.89	1.96
chrY	0.50	1.90	10.83	1.05	1.47	1.16	4.67
Whole genome	0.44	0.61	5.53	1.04	1.22	0.81	1.81

Non-B DNA motifs located at repeats and satellites are usually underrepresented but occasionally are enriched. It has been shown previously, via the analysis of non-T2T genomes, that some types of non-B DNA motifs are present at repeats, and might be propagated via the spreading of repeats [53,68,69]. The T2T genomes provided a complete resolution of repeats in the ape genomes, including transposable elements and satellites. We found that non-B DNA motifs to be underrepresented for most repeat and satellite types (Fig. 5). However, they were occasionally enriched, especially at satellites (see Fig. 5 and Fig. S3A-B for human and Fig. S2A-F for the remaining ape species). The patterns were often shared across species, for example the satellites LSAU and MSR1 were enriched for G-quadruplexes in all the species included in this study. The satellites Nereid, Neso and Proteus were enriched for direct repeats, inverted repeats and mirrored repeats in all the studied non-human apes. The human repeat annotation has been updated with manually curated satellites and composite repeats (Fig. S3A-B). Many of these satellites were enriched for direct repeats, but we found no enrichment of G-quadruplex motifs in Walusat (Fig. S3A), which was reported previously [70]. For composite repeats, G-quadruplex and Z-DNA motifs were more often enriched than other non-B types.

We further analyzed some of the human satellites and repeats in which G-quadruplex motifs were enriched (Table 4). These included retrotransposon SVA, and satellites ACRO1, GSAT, GSATII, GSATX, LSAU, SST1, TAR1, HSAT5, and MSR1. Whereas G-quadruplex motifs were enriched in these cases, we had no evidence of G-quadruplex formation or non-formation. As a proxy of such formation, we evaluated methylation profiles for G-quadruplexes and repeats harboring them, as methylation was shown to be antagonistic to G4 formation [71,72]. Using the methylome of the HG002 lymphoblastoid cell line [73], we discovered that G4 motifs enriched at repeats and satellites were methylated in most cases, reflecting the overall methylation trends for these genomic regions (Fig. 6). This was particularly evident for G4s located in the SVA retrotransposon. However, for certain satellites (e.g., LSAU and TAR1), the methylation level for G4s was lower than that for the satellites harboring them. For the HSAT5 satellites, we observed a bimodal distribution of methylated vs. unmethylated G4s, suggesting that some of them form alternative structures.

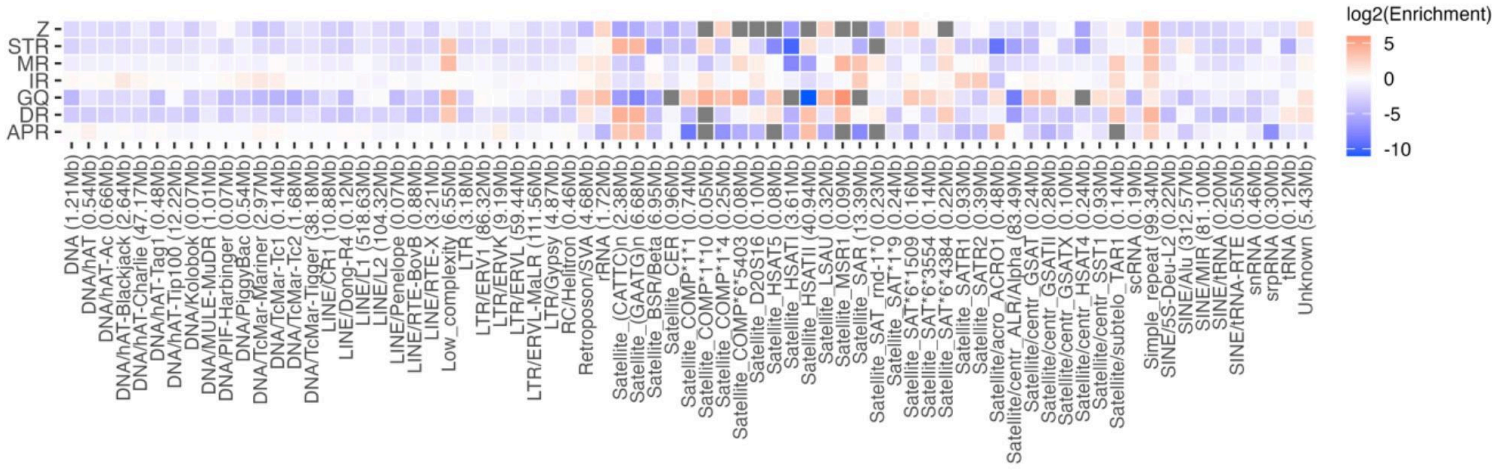

Figure 5. Enrichment of non-B DNA at repeats, given as log-fold densities compared to genome-wide densities for the human genome. Underrepresentation of non-B DNA (values below 1) are shown in blue, whereas enrichment (values above 1) are shown in red. Long repeat names have been shortened for visualization purposes (marked with *). Total repeat lengths are given after the names, repeats with a total length shorter than 50 kb are not shown.

Table 4. Repeats and satellites enriched for G-quadruplex motifs, and their average methylation scores, in the human genome.

Repeat	Average methylation in G4 motifs	Number of CpG sites in G4s	Average methylation in full repeats	Number of CpG sites in full repeats
Retroposon/SVA	0.89	9,600	0.91	234,912
rRNA	0.52	8,856	0.57	73,832
Satellite/acro_ACRO1	0.86	299	0.59	10,241
Satellite/centr_GSAT	0.57	616	0.63	5,898
Satellite/centr_GSATII	0.59	1,169	0.63	10,920
Satellite/centr_GSATX	0.80	64	0.76	3,610
Satellite/centr_SST1	0.72	1,285	0.78	44,309
Satellite/subtelo_TAR1	0.61	442	0.75	13,887
Satellite_COMP-subunit_LSAU-BSAT_rnd-1_family-1	0.77	2,931	0.78	46,344
Satellite_COMP-subunit_LSAU-BSAT_rnd-1_family-4	0.64	1,506	0.73	13,009
Satellite_COMP-subunit_LSAU-BSAT_rnd-6_family-5403	0.63	1,156	0.69	6,603
Satellite_HSAT5	0.52	47	0.51	710
Satellite_LSAU	0.65	2,424	0.76	20,211
Satellite_MSR1	0.47	450	0.44	821
Satellite_SAT-VAR_rnd-6_family-3554	0.52	53	0.54	1,266
Genome-wide average	**0.44**	**723,187**	**0.67**	**33,328,344**

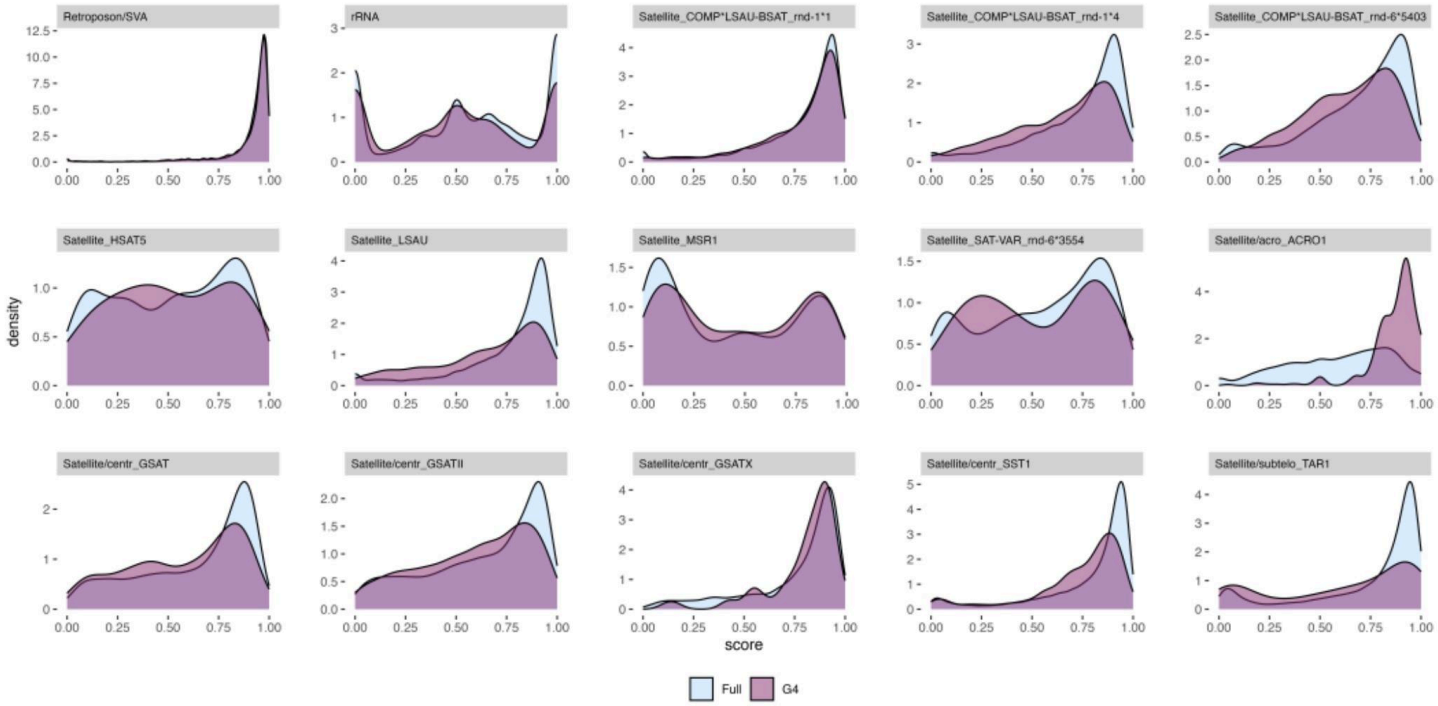

Figure 6. Distribution of methylation scores at different repeats (light blue) and G-quadruplexes (purple) overlapping with the repeat annotations, obtained from CpG sites in the human lymphoblastoid cell line (HG002). A score of 1 means all reads were methylated for this site, whereas 0 means no methylation.

Discussion

We conducted a detailed analysis of non-B DNA motifs in the T2T assemblies of the genomes of human and non-human apes, which have recently become available [60–63]. We have found an overrepresentation of most types of non-B DNA motifs in the newly added sequences of the human T2T genome, in agreement with previous studies of non-human T2T ape sex chromosomes [62] and autosomes [63]. Therefore, we were able to analyze previously inaccessible regions of the genome that are rich in non-B DNA motifs and thus uncovered the complete genome-wide repertoire of such motifs across human and other available non-human primate T2T genomes. Thus, our study complements previous studies of non-B DNA motif enrichment in previously sequenced regions of the human genome (e.g., [74]).

We found that, on a large scale, non-B DNA motifs were distributed unevenly among and along ape genomes. In the human genome, many blocks of high-density non-B DNA motifs corresponded either to centromeric satellites or to satellites at the short arms of acrocentric chromosomes. Centromeric satellites (see Fig. 1 in [60]) corresponded to A-phased and direct repeats. In contrast, short arms of acrocentric chromosomes (see Fig. 4 in [60]) corresponded to a patchwork of different non-B DNA motif type combinations. For instance, the Hsat1 satellite was rich in inverted and mirror repeats and Hsat3 was rich in A-phased and direct repeats and in STRs. The ability of these satellites to form alternative DNA structures might play an important role in their copy number dynamics and in determining their self-similarity. Therefore, these non-B DNA features should be incorporated into future models of satellite evolution.

Our examination of active, experimentally defined centromeres in great apes indicated that most of them were enriched in at least one type of non-B DNA motifs. This extends an earlier study of non-B DNA enrichment at active centromeres of human, African monkey, and mouse [20] to complete chromosome sequences of multiple species of great apes and suggests an important role of non-B DNA structures in defining centromeres.

We have identified several instances of non-B DNA motif enrichment at particular satellites and transposable elements. For G4 motifs in particular, we were able to predict their formation at such regions by assigning their methylation status, as methylation inhibits G4 formation. In many instances, G4s enriched at transposable

elements and satellites were methylated and thus are unlikely to form. However, in some instances, G4s were less methylated than the overall satellites they are embedded into, had a bimodal methylation density distribution, or were undermethylated. Such G4s should be investigated further as they may have functional significance for the satellites. Moreover, the methylomes of additional cell lines should be added to this analysis, as, for instance, satellites in the CHM13 cell lines are known to be usually unmethylated [73].

Additional, more direct experimental studies should be performed to investigate the formation of non-B DNA structures in ape cells and tissues. Such experiments should also elucidate the precise structure these motifs are forming, particularly when the same sequence is being annotated as multiple non-B DNA motif types. Distinguishing these structures can be important, as, for instance, the promoter of human C-MYC oncogene can form either a G-quadruplex or H-DNA, which might have different effects on genomic instability in this region [31]. Similarly, knowing what particular non-B DNA structures form at satellites can inform their expansion mode.

Methods

Non-B DNA annotation. Non-B DNA motifs were annotated as described in [63]. In short, motifs of A-phased repeats, direct repeats, mirror repeats, short tandem repeats and Z-DNA were annotated in each T2T genome with the software gfa (https://github.com/abcsFrederick/non-B_gfa) with the flag *–skipGQ*. G-quadruplexes were annotated using Quadron [67] with default settings. The output from each motif type was converted to bedformat and any overlapping annotations of the same type were merged with mergeBed from bedtools v 2.31.1 [75]. For G-quadruplexes, motifs without scores were omitted from the analysis. Overlap between different motif types was retrieved using bedtools and in-house scripts (see github link below). For spacer length analysis, the spacers were extracted from the raw gfa output files.

Alignments to old assembly versions. Each of the T2T assemblies (CHM13v2.0 and v.2 assemblies for non-human ape genomes as available in [63]) for which there were an older non-T2T genome available, were mapped to its older counterpart (panPan3 for bonobo, panTro6 for chimpanzee, hg38 for human, gorGor6 for gorilla and ponAbe3 for Sumatran orangutan) using winnowmap v2.03 [76]. We followed the winnowmap recommendations and first generated a set of high-frequency *k*-mers with meryl v1.4.1 [77] using *k*=19. Regions that did not map to the old assembly were extracted using bedtools complement, and assigned as 'new' (note that newly added regions that are duplicates of previously assembled sequence, e.g. previously unresolved multi-copy genes, repetitive arrays. etc., can align in a many-to-one fashion and will then not be considered as new). Densities of non-B motifs in 'new' and 'old' sequences (i.e. sequence in T2T genomes that did not align vs aligned to the older assembly versions) were extracted with bash and awk scripts, and fold enrichment was calculated as density in 'new' divided by density in 'old'. The number of non-B annotated base pairs in new and old sequence were compared in a chi-square goodness of fit test for each non-B motif type and chromosome type separately, corrected for multiple testing with the Bonferroni method.

Non-B distribution along the chromosomes and enrichment at centromeres. Non-B motif density along the genome was calculated for 100-kb windows, generated with bedtools makewindows. For visualization purposes, the densities of each non-B motif type were normalized by the highest value to a scale from 0 to 1. Centromeric regions were taken from the GenomeFeatures tracks downloaded from https://www.genomeark.org/ and converted to bed format. For chromosomes with two annotated 'CEN' regions with a 'SAT' in between, we merged them to a single region for visualization purposes, however only CEN regions were used for the enrichment analysis. Fold enrichment was calculated as non-B motif density within centromeres compared to genome-wide non-B density.

Enrichment upstream of TSSs. Gene annotations for human (CHM13v2.0) were downloaded from https://s3-us-west-2.amazonaws.com/human-pangenomics/T2T/CHM13/assemblies/annotation/chm13v2.0_RefSeq_Liftoff_v5.2.gff3.gz. Only transcripts annotated as 'mRNA' were used to extract regions of 1 kb upstream from the start position, and these regions were intersected with non-B DNA motifs regardless of strand. Fold enrichment was calculated in the same manner as described above for centromeres. For statistical testing, we randomly generated 1000 bedfiles from the human genome using bedtools random, having the same number of sequences (31,708) and lengths (1kb) as the real TSS file. These random regions were intersected with nonB DNA annotation and fold enrichment was calculated and used as a random distribution to which the TSS enrichment was compared to, using a two-sided permutation test with 10,000 permutations.

Enrichment at repetitive sequences and methylation analysis. RepeatMasker annotations were downloaded from https://github.com/marbl/CHM13 (for human version CHM13v2.0) and from https://www.genomeark.org/t2t-all/ (for all other apes). For human, also manually curated repeat annotations of new satellites and composite repeats were downloaded from the same source and analyzed separately. RepeatMasker output was converted to bedformat and labeled according to the repeat class for all repeats except satellites, where both the class and the specific names were used. The repeats were intersected with each non-B motif type separately using bedtools, and non-B density in each repeat class were compared to the genome-wide density using in-house python and bash scripts (see github). Methylation data from the H002 cell line, translated into CHM13v2.0 coordinates, was downloaded from https://s3-us-west-2.amazonaws.com/human-pangenomics/T2T/CHM13/assemblies/annotation/regulation/chm13v2.0_hg002_CpG_ont_guppy6.1.2.bed [73]. This file contains methylation scores for CpG sites, given as a fraction of methylated reads (where 0 means no methylation was detected, and 1 means that all reads were

methylated). We extracted G-quadruplexes overlapping (partially or fully) with repeat classes that had shown an enrichment for G4s in the above analysis, and compared the distribution of methylation scores within the G-quadruplex motifs with the distribution of methylation scores within the full repeats of each repeat class.

All figures were generated in R v4.4.0 using the tidyverse library [78].

Code availability

All code and in-house scripts generated for this paper are available on github:
https://github.com/makovalab-psu/T2T_primate_nonB

Acknowledgments

We are grateful to Saswat Mohanty, Edmundo Torres-Gonzalez, Karol Pal, Francesca Chiaromonte, and Jacob Sieg for discussions of the results and useful suggestions. Robert Harris provided code for running winnowmap. Eddie Yong Hwee Loh provided useful explanations for methylation analysis. Matthias Weissensteiner helped with generating Figure 1.

References

1. Watson, J. D. & Crick, F. H. C. Genetical Implications of the Structure of Deoxyribonucleic Acid. Nature 171, 964–967 (1953).

2. Guiblet, W. M. et al. Long-read sequencing technology indicates genome-wide effects of non-B DNA on polymerization speed and error rate. Genome Res. 28, 1767–1778 (2018).

3. Saranathan, N. & Vivekanandan, P. G-Quadruplexes: More Than Just a Kink in Microbial Genomes. Trends Microbiol. 27, 148–163 (2019).

4. Marsico, G. et al. Whole genome experimental maps of DNA G-quadruplexes in multiple species. Nucleic Acids Res. 47, 3862–3874 (2019).

5. Gazanion, E. et al. Genome wide distribution of G-quadruplexes and their impact on gene expression in malaria parasites. PLoS Genet. 16, e1008917 (2020).

6. Puig Lombardi, E. et al. Thermodynamically stable and genetically unstable G-quadruplexes are depleted in genomes across species. Nucleic Acids Res. 47, 6098–6113 (2019).

7. Brázda, V. et al. G-Quadruplexes in the Archaea Domain. Biomolecules 10, (2020).

8. Kouzine, F. et al. Permanganate/S1 Nuclease Footprinting Reveals Non-B DNA Structures with Regulatory Potential across a Mammalian Genome. Cell Syst 4, 344–356.e7 (2017).

9. Poggi, L. & Richard, G.-F. Alternative DNA Structures In Vivo : Molecular Evidence and Remaining Questions. Microbiology and Molecular Biology Reviews vol. 85 Preprint at https://doi.org/10.1128/mmbr.00110-20 (2021).

10. Wang, G. & Vasquez, K. M. Dynamic alternative DNA structures in biology and disease. Nat. Rev. Genet. 24, 211–234 (2023).

11. Matos-Rodrigues, G., Hisey, J. A., Nussenzweig, A. & Mirkin, S. M. Detection of alternative DNA structures and its implications for human disease. Mol. Cell 83, 3622–3641 (2023).

12. Zyner, K. G. et al. G-quadruplex DNA structures in human stem cells and differentiation. Nat. Commun. 13, 142 (2022).

13. Matos-Rodrigues, G. et al. S1-END-seq reveals DNA secondary structures in human cells. Mol. Cell (2022) doi:10.1016/j.molcel.2022.08.007.

14. Fleming, A. M. & Burrows, C. J. Interplay of Guanine Oxidation and G-Quadruplex Folding in Gene Promoters. J. Am. Chem. Soc. 142, 1115–1136 (2020).

15. Roychoudhury, S. et al. Endogenous oxidized DNA bases and APE1 regulate the formation of G-quadruplex structures in the genome. Proc. Natl. Acad. Sci. U. S. A. 117, 11409–11420 (2020).

16. Selvam, S., Koirala, D., Yu, Z. & Mao, H. Quantification of topological coupling between DNA superhelicity and G-quadruplex formation. J. Am. Chem. Soc. 136, 13967–13970 (2014).

17. Prorok, P. et al. Involvement of G-quadruplex regions in mammalian replication origin activity. Nat. Commun. 10, 3274 (2019).

18. Akerman, I. et al. A predictable conserved DNA base composition signature defines human core DNA replication origins. Nat. Commun. 11, 4826 (2020).

19. Sahakyan, A. B., Murat, P., Mayer, C. & Balasubramanian, S. G-quadruplex structures within the 3' UTR of LINE-1 elements stimulate retrotransposition. Nat. Struct. Mol. Biol. 24, 243–247 (2017).

20. Kasinathan, S. & Henikoff, S. Non-B-Form DNA Is Enriched at Centromeres. Mol. Biol. Evol. 35, 949–962 (2018).

21. Moye, A. L. et al. Telomeric G-quadruplexes are a substrate and site of localization for human telomerase. Nat. Commun. 6, 7643 (2015).

22. Haran, T. E. & Mohanty, U. The unique structure of A-tracts and intrinsic DNA bending. Quarterly Reviews of Biophysics vol. 42 41–81 Preprint at https://doi.org/10.1017/s0033583509004752 (2009).

23. Spiegel, J. et al. G-quadruplexes are transcription factor binding hubs in human chromatin. Genome Biol. 22, 117 (2021).

24. Gong, J.-Y. et al. G-quadruplex structural variations in human genome associated with single-nucleotide variations and their impact on gene activity. Proc. Natl. Acad. Sci. U. S. A. 118, (2021).

25. Gazanion, E. et al. Genome wide distribution of G-quadruplexes and their impact on gene expression in malaria parasites. PLoS Genet. 16, e1008917 (2020).

26. Biswas, B., Kandpal, M. & Vivekanandan, P. A G-quadruplex motif in an envelope gene promoter regulates transcription and virion secretion in HBV genotype B. Nucleic Acids Research vol. 45 11268–11280 Preprint at https://doi.org/10.1093/nar/gkx823 (2017).

27. Shin, S.-I. et al. Z-DNA-forming sites identified by ChIP-Seq are associated with actively transcribed regions in the human genome. DNA Res. 23, 477–486 (2016).

28. Sulovari, A. et al. Human-specific tandem repeat expansion and differential gene expression during primate evolution. Proc. Natl. Acad. Sci. U. S. A. 116, 23243–23253 (2019).

29. Roberts, J. W. Mechanisms of Bacterial Transcription Termination. Journal of Molecular Biology vol. 431 4030–4039 Preprint at https://doi.org/10.1016/j.jmb.2019.04.003 (2019).

30. Yamamoto, Y., Miura, O. & Ohyama, T. Cruciform Formable Sequences within Pou5f1 Enhancer Are Indispensable for Mouse ES Cell Integrity. International Journal of Molecular Sciences vol. 22 3399 Preprint at https://doi.org/10.3390/ijms22073399 (2021).

31. Del Mundo, I. M. A., Zewail-Foote, M., Kerwin, S. M. & Vasquez, K. M. Alternative DNA structure

formation in the mutagenic human c-MYC promoter. Nucleic Acids Res. 45, 4929–4943 (2017).

32. Georgakopoulos-Soares, I. et al. High-throughput characterization of the role of non-B DNA motifs on promoter function. Cell Genomics vol. 2 100111 Preprint at https://doi.org/10.1016/j.xgen.2022.100111 (2022).

33. Roy, S. S. et al. Artificially inserted strong promoter containing multiple G-quadruplexes induces long-range chromatin modification. Elife 13, (2024).

34. Hänsel-Hertsch, R. et al. G-quadruplex structures mark human regulatory chromatin. Nat. Genet. 48, 1267–1272 (2016).

35. Lago, S. et al. Promoter G-quadruplexes and transcription factors cooperate to shape the cell type-specific transcriptome. Nat. Commun. 12, 3885 (2021).

36. Miura, O., Ogake, T., Yoneyama, H., Kikuchi, Y. & Ohyama, T. A strong structural correlation between short inverted repeat sequences and the polyadenylation signal in yeast and nucleosome exclusion by these inverted repeats. Current Genetics vol. 65 575–590 Preprint at https://doi.org/10.1007/s00294-018-0907-8 (2019).

37. Hou, Y. et al. Integrative characterization of G-Quadruplexes in the three-dimensional chromatin structure. Epigenetics 14, 894–911 (2019).

38. Robinson, J., Raguseo, F., Nuccio, S. P., Liano, D. & Di Antonio, M. DNA G-quadruplex structures: more than simple roadblocks to transcription? Nucleic Acids Res. 49, 8419–8431 (2021).

39. Georgakopoulos-Soares, I. et al. Alternative splicing modulation by G-quadruplexes. Nat. Commun. 13, 2404 (2022).

40. Bochman, M. L., Paeschke, K. & Zakian, V. A. DNA secondary structures: stability and function of G-quadruplex structures. Nat. Rev. Genet. 13, 770–780 (2012).

41. Bugaut, A. & Balasubramanian, S. 5'-UTR RNA G-quadruplexes: translation regulation and targeting. Nucleic Acids Res. 40, 4727–4741 (2012).

42. Lyu, K., Chow, E. Y.-C., Mou, X., Chan, T.-F. & Kwok, C. K. RNA G-quadruplexes (rG4s): genomics and biological functions. Nucleic Acids Res. (2021) doi:10.1093/nar/gkab187.

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

44. Wang, G. & Vasquez, K. M. Impact of alternative DNA structures on DNA damage, DNA repair, and genetic instability. DNA Repair 19, 143–151 (2014).

45. Kaushal, S. & Freudenreich, C. H. The role of fork stalling and DNA structures in causing chromosome fragility. Genes Chromosomes Cancer 58, 270–283 (2019).

46. Twayana, S. et al. Translesion polymerase eta both facilitates DNA replication and promotes increased human genetic variation at common fragile sites. Proc. Natl. Acad. Sci. U. S. A. 118, (2021).

47. Bournique, E., Dall'Osto, M., Hoffmann, J.-S. & Bergoglio, V. Role of specialized DNA polymerases in the limitation of replicative stress and DNA damage transmission. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis vol. 808 62–73 Preprint at https://doi.org/10.1016/j.mrfmmm.2017.08.002 (2018).

48. Tsao, W.-C. & Eckert, K. A. Detours to Replication: Functions of Specialized DNA Polymerases during Oncogene-induced Replication Stress. Int. J. Mol. Sci. 19, (2018).

49. Boyer, A.-S., Grgurevic, S., Cazaux, C. & Hoffmann, J.-S. The human specialized DNA polymerases and non-B DNA: vital relationships to preserve genome integrity. J. Mol. Biol. 425, 4767–4781 (2013).

50. Sauer, M. & Paeschke, K. G-quadruplex unwinding helicases and their function. Biochem. Soc. Trans. 45, 1173–1182 (2017).

51. McKinney, J. A., Wang, G. & Vasquez, K. M. Distinct mechanisms of mutagenic processing of alternative DNA structures by repair proteins. Mol Cell Oncol 7, 1743807 (2020).

52. McGinty, R. J. & Sunyaev, S. R. Mutagenesis at non-B DNA motifs in the human genome: a course correction. Preprint at https://doi.org/10.1101/2022.02.08.479604.

53. Makova, K. D. & Weissensteiner, M. H. Noncanonical DNA structures are drivers of genome evolution. Trends Genet. 39, 109–124 (2023).

54. Haeusler, A. R. et al. C9orf72 nucleotide repeat structures initiate molecular cascades of disease. Nature 507, 195–200 (2014).

55. Tateishi-Karimata, H. & Sugimoto, N. Roles of non-canonical structures of nucleic acids in cancer and neurodegenerative diseases. Nucleic Acids Res. 49, 7839–7855 (2021).

56. Cheloshkina, K. & Poptsova, M. Comprehensive analysis of cancer breakpoints reveals signatures of genetic and epigenetic contribution to cancer genome rearrangements. PLoS Comput. Biol. 17, e1008749 (2021).

57. Maizels, N. G4-associated human diseases. EMBO Rep. 16, 910–922 (2015).

58. Weissensteiner, M. H. et al. Accurate sequencing of DNA motifs able to form alternative (non-B) structures. Genome Res. 33, 907–922 (2023).

59. McGinty, R. J. & Sunyaev, S. R. Revisiting mutagenesis at non-B DNA motifs in the human genome. Nat. Struct. Mol. Biol. 30, 417–424 (2023).

60. Nurk, S. et al. The complete sequence of a human genome. Science 376, 44–53 (2022).

61. Rhie, A. et al. The complete sequence of a human Y chromosome. Nature 621, 344–354 (2023).

62. Makova, K. D. et al. The complete sequence and comparative analysis of ape sex chromosomes. Nature 630, 401–411 (2024).

63. Yoo, D. et al. Complete sequencing of ape genomes. bioRxiv (2024) doi:10.1101/2024.07.31.605654.

64. Cox, R. & Mirkin, S. M. Characteristic enrichment of DNA repeats in different genomes. Proc. Natl. Acad. Sci. U. S. A. 94, 5237–5242 (1997).

65. Zhao, J., Bacolla, A., Wang, G. & Vasquez, K. M. Non-B DNA structure-induced genetic instability and evolution. Cell. Mol. Life Sci. 67, 43–62 (2010).

66. Sinden, R. R., Zheng, G. X., Brankamp, R. G. & Allen, K. N. On the deletion of inverted repeated DNA in Escherichia coli: effects of length, thermal stability, and cruciform formation in vivo. Genetics 129, 991–1005 (1991).

67. Sahakyan, A. B. et al. Machine learning model for sequence-driven DNA G-quadruplex formation. Sci. Rep. 7, 14535 (2017).

68. Kejnovsky, E., Tokan, V. & Lexa, M. Transposable elements and G-quadruplexes. Chromosome Res. 23, 615–623 (2015).

69. Kejnovsky, E. & Lexa, M. Quadruplex-forming DNA sequences spread by retrotransposons may serve as genome regulators. Mob. Genet. Elements 4, e28084 (2014).

70. Hoyt, S. J. et al. From telomere to telomere: The transcriptional and epigenetic state of human repeat elements. Science 376, eabk3112 (2022).

71. Halder, R. et al. Guanine quadruplex DNA structure restricts methylation of CpG dinucleotides genome-wide. Mol. Biosyst. 6, 2439–2447 (2010).

72. Mao, S.-Q. et al. DNA G-quadruplex structures mold the DNA methylome. Nat. Struct. Mol. Biol. 25, 951–957 (2018).

73. Gershman, A. et al. Epigenetic patterns in a complete human genome. Science 376, eabj5089 (2022).

74. Guiblet, W. M. et al. Selection and thermostability suggest G-quadruplexes are novel functional elements of the human genome. Genome Res. 31, 1136–1149 (2021).

75. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842 (2010).

76. Jain, C. et al. Weighted minimizer sampling improves long read mapping. Bioinformatics 36, i111–i118 (2020).

77. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biol. 21, 245 (2020).

78. Wickham, H. et al. Welcome to the Tidyverse. Journal of Open Source Software 4, 1686 (2019).