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

Linnéa Smeds[1], *Kaivan Kamali*[1], *Iva Kejnovská*[2], Eduard Kejnovský[3], Francesca Chiaromonte[4,5,6], *Kateryna D. Makova*[1,5]

[1]Department of Biology, Penn State University, University Park, PA 16802, USA

[2]Department of Biophysics of Nucleic Acids, Institute of Biophysics of the Czech Academy of Sciences, Královopolská 135, 612 65 Brno, Czech Republic

[3]Department of Plant Developmental Genetics, Institute of Biophysics of the Czech Academy of Sciences, Královopolská 135, 612 65 Brno, Czech Republic

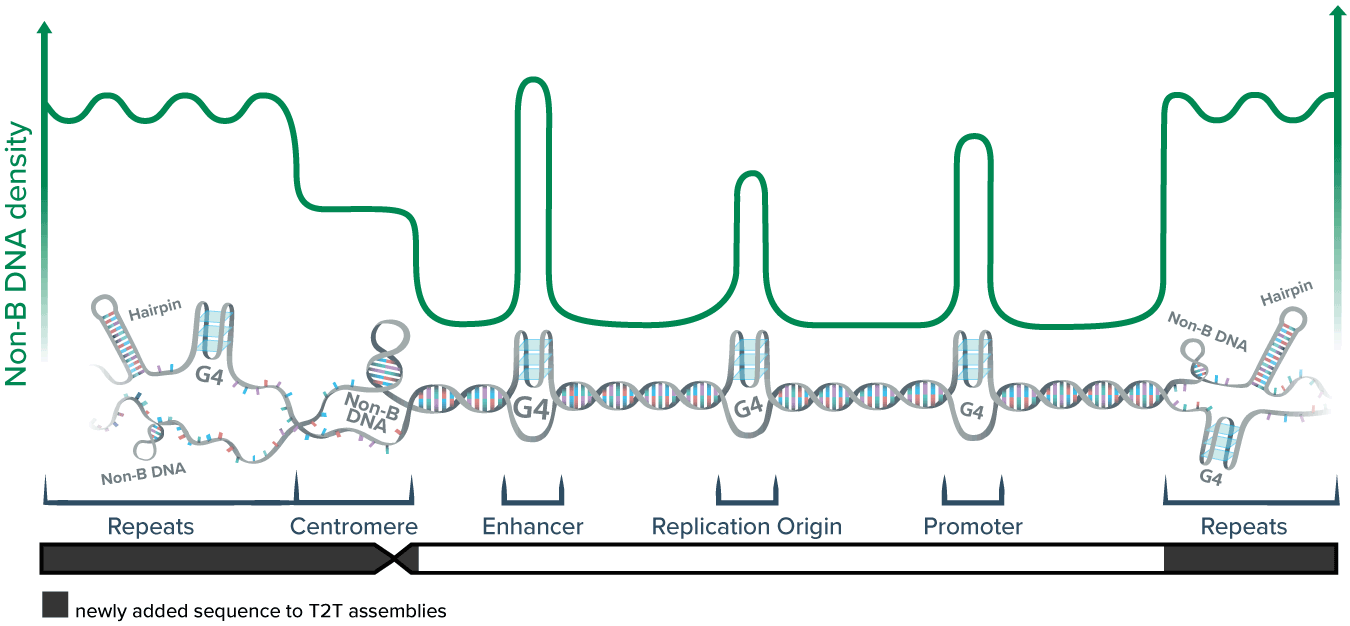
[4]Department of Statistics, Penn State University, University Park, PA 16802, USA

[5]Center for Medical Genomics, Penn State University, University Park, PA 16802 USA

[6]L'EMbeDS, Sant'Anna School of Advanced Studies, 56127 Pisa, Italy

Correspondence to Kateryna Makova (kdm16@psu.edu)

Graphical Abstract



Abstract

Non-canonical (non-B) DNA structures—e.g., bent DNA, hairpins, G-quadruplexes, Z-DNA, etc.—which form at certain sequence motifs (e.g., A-phased repeats, inverted repeats, etc.), have emerged as important regulators of cellular processes and drivers of genome evolution. Yet, they have been understudied due to their repetitive nature and potentially inaccurate sequences generated with short-read technologies. Here we comprehensively characterize such motifs in the long-read telomere-to-telomere (T2T) genomes of human, bonobo, chimpanzee, gorilla, Bornean orangutan, Sumatran orangutan, and siamang. Non-B DNA motifs are enriched at the genomic regions added to T2T assemblies, and occupy 9-15%, 9-11%, and 12-38% of autosomes, and chromosomes X and Y, respectively. Functional regions (e.g., promoters and enhancers) and repetitive sequences are enriched in non-B DNA motifs. Non-B DNA motifs concentrate at short arms of acrocentric chromosomes in a pattern reflecting their satellite repeat content and might contribute to satellite dynamics in these regions. Most centromeres and/or their flanking regions are enriched in at least one non-B DNA motif type, consistent with a potential 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 previously inaccessible genomic regions.

Introduction

In addition to canonical B DNA—the right-handed double helix with 10 base pairs per turn[1]—an estimated 13% of the human genome has the ability 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 (or H-DNA) formed by some 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 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: DNA at a non-B motif can fold into either B or non-B form at a given time. For instance, folding into non-B forms is sensitive to oxidative stress[3,4] and to temporal signals during cell differentiation and development[5,6].

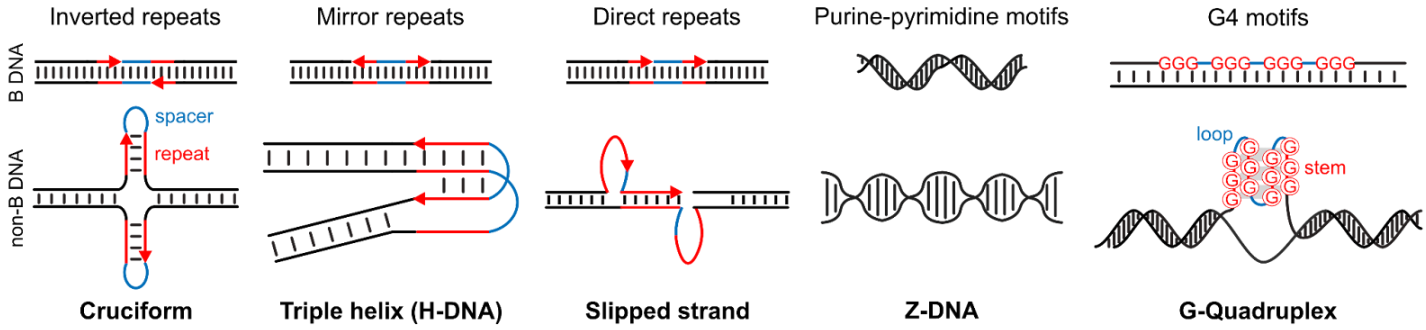


Figure 1. Types of non-B DNA. For structures formed from repeats: repeats are red, spacers are blue. For G4s: stems are red, loops are blue. Triple helix is formed by mirror repeats comprising predominantly purines or pyrimidines and having a short spacer (less than 9 bp)[7].

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[8,9]. G4s affect the life cycle of L1 transposable elements[10] and protect chromosome ends at telomeres[11]. Non-B DNA has been implicated in regulating transcription[12–24]. G4s regulate chromatin organization[24–29] and methylation of CpG islands[30]. The transcribed non-B DNA motifs can form structured RNA, which regulates alternative splicing[31], translation of mRNA[16,32,33], and function of non-coding RNA[34].

Non-B DNA has also been implicated in the definition and function of centromeres. For example, inverted repeats forming non-B DNA have been hypothesized to define centromeres[35], which would resolve the CENP-B paradox. Indeed, non-B DNA might play a role attributed to CENP-B, the highly conserved protein binding motif present in centromeres across a range of taxa and proposed to be involved in centromere formation—but paradoxically missing entirely on some chromosomes[36,37]. Recent studies also found enrichment at centromeres for G4s in Drosophila[38] and Z-DNA, and A-phased, direct, and mirror repeats in plants[39,40] and argued that non-B DNA is important for centromere activity and stability.

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*[41–43]. The cell recruits error-prone specialized helicases[44] and polymerases[45–48] to handle non-B DNA structures during replication. Moreover, non-B DNA affects the efficiency of DNA repair pathways[30,42,49]. Increased mutagenesis and genomic instability at non-B DNA are evident in cancers with mutated components of these pathways[42]. In non-cancerous cells, the effects of non-B DNA on replication progression, mutation rate, and genome instability remain controversial[50]. Nevertheless, non-B DNA has been recognized as an important driver of genome evolution[51].

Non-B DNA structures have been implicated in neurodegenerative diseases (e.g., amyotrophic lateral sclerosis[52] and fragile X syndrome[53]). They are also the preferential sites of genome rearrangements[54] and affect gene expression[53] in cancers. Some diseases result from mutations in genes encoding proteins

processing non-B DNA (e.g., Werner syndrome)[55].

Despite its unequivocal importance for genome function, mutations, and diseases, studying non-B DNA has been challenging for several reasons. First, several sequencing technologies, and in particular short-read Illumina technology, have increased error rates at non-B DNA motifs[56,57]. To overcome this limitation, the current recommendation is to use multiple long-read sequencing technologies as they differ in their biases at non-B DNA motifs[56]. Second, incomplete genome assemblies have hindered the full characterization of non-B DNA motifs, particularly for the ones located at repetitive regions.

Here, we identify non-B DNA motif occurrences in the complete, telomere-to-telomere (T2T) genomes of human[58,59] and several non-human apes—bonobo and chimpanzee (which diverged from each other ~2.5 million years ago, Mya, and from the human lineage 7 Mya), gorilla (which diverged from human and bonobo/chimpanzee 9 Mya), Bornean and Sumatran orangutans (which diverged from each other ~1 Mya, and from the previously mentioned species 17 Mya), and the lesser ape siamang (which diverged from great apes 20 Mya)[60,61]. This provides a comprehensive view of non-B DNA genomic distribution across most living great ape species and an outgroup. Importantly, the human and primate T2T assemblies employed in our study were produced with two long-read sequencing technologies, thus minimizing the effects of sequencing biases at non-B DNA motifs. Using this exhaustive dataset, we tackle several questions that could not be addressed prior to the availability of complete ape genomes, including the potential enrichment of non-B DNA at centromeres and satellites. We further investigate G4 formation in satellites using methylation data from two cell lines, and validate some commonly found G4 motifs experimentally with circular dichroism (CD) analysis.

Results

Non-B DNA annotations. Most non-B DNA motifs—A-phased, direct, inverted, and mirror repeats, short tandem repeats (STRs), and Z-DNA – were annotated in the latest versions of human and non-human ape T2T genomes[61] with gfa[7,61]. Some previous studies suggested that non-B DNA folds at inverted repeats only when the spacer length is below 15 bp[62–64], 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 most of our annotations had spacers below 15 bp (Fig. S1). We have also annotated a subset of mirror repeats with a high potential to form triplexes (i.e. mirror repeats comprising predominantly purines or pyrimidines and having a short spacer, see Methods) also using gfa. G4s were annotated with Quadron[65].

An overrepresentation of non-B DNA motifs in the newly added regions of the human T2T genomes. We observed an overrepresentation of most non-B DNA motif types in the newly added 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, and mirror repeats, as well as STRs, were substantially overrepresented at the newly added sequences for the autosomes. Direct, inverted, and mirror repeats, as well as G4s (although this was not significant after correcting for multiple tests) and STRs, were overrepresented at such sequences for the X chromosome, and inverted and mirror repeats were overrepresented for the Y chromosome. Similar results were previously obtained for great ape sex chromosomes[59,60] and autosomes[61].

Table 1. Non-B DNA motifs enriched at the newly added sequences of the human T2T genome as compared to hg38, shown as fold enrichment for unaligned vs. aligned sequences. Bold numbers indicate significantly different non-B DNA content in aligned vs. unaligned sequences (chi-square goodness of fit test with Bonferroni correction for multiple testing, *P*<0.01. Cells marked with '*" remained significant after the data had been randomly subsampled down to half in 10 independent runs, see Methods). APR: A-phased repeats, DR: direct repeats, G4: G-quadruplexes, IR: inverted repeats, MR: mirror repeats, STR: short tandem repeats.

Chr/Non-B type	APR	DR	STR	IR	MR	G4	Z-DNA	All Non-B
Autosomes	**6.72***	**18.75***	**13.44***	**1.63***	**2.61***	**0.96***	**0.65***	**4.98***
Chr X	**0.55**	**24.92***	**6.94***	**7.66***	**13.83***	**2.76**	0.98	**5.95***
Chr Y	**0.08***	**0.82***	**0.44***	**19.70***	**19.43***	**0.08***	**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 (Fig. 2 and Table S1). Depending on species, autosomes had 9.2-14.9% of their sequence annotated in non-B DNA motifs. The X chromosome had a lower percentage, and the Y chromosome had a higher percentage than that for the autosomes in each species (ranging across species from 8.8-11.4% for the X and from 12.2-37.9% for the Y). As a rule, inverted, mirror, direct repeats, and STRs were more abundant than the other non-B DNA motif types. G4s occupied 0.8-1.0% of autosomal sequences and usually a lower percentage of sex chromosomal sequences. A-phased repeats, triplex motifs (Table S1B), and Z-DNA each occupied a lower percentage of autosomal sequences than G4s. Among the species analyzed, the gorilla genome had the highest percentage of non-B DNA motif annotations, whereas siamang had the lowest. Some non-B DNA motif types were distinctly more abundant in some species than in others. For instance, across chromosome types, direct repeats were more abundant in gorilla than in other species.

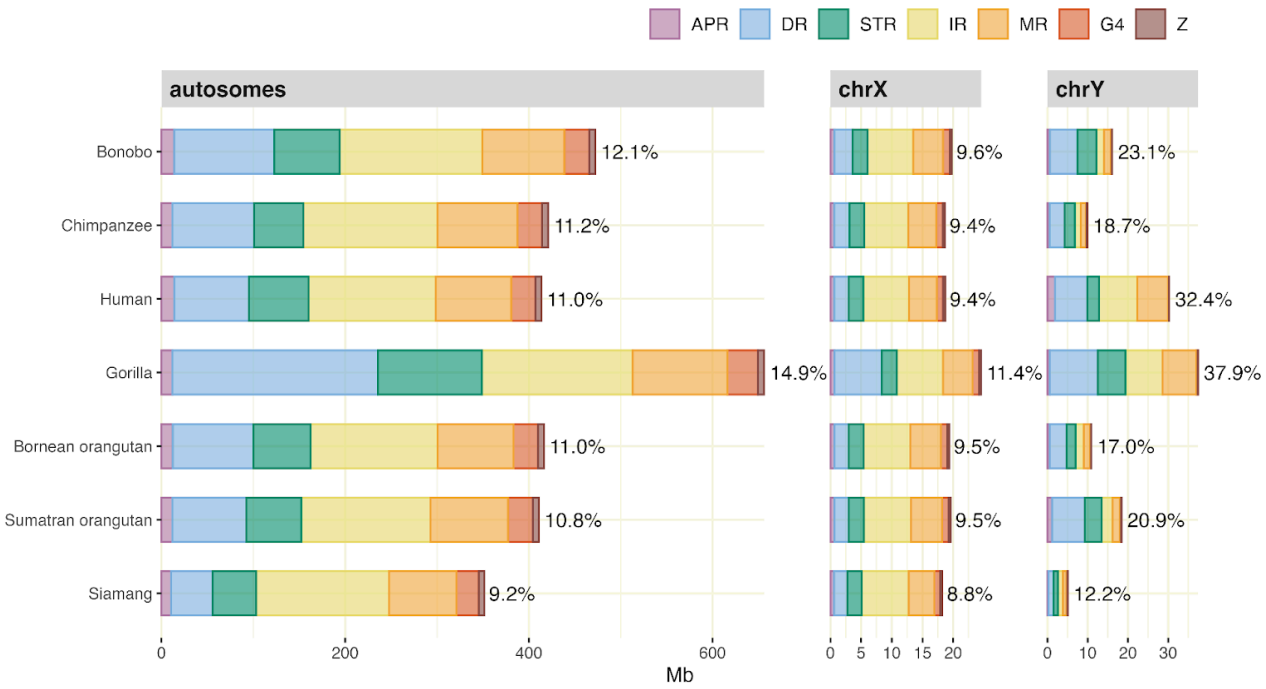


Figure 2. Non-B DNA motif annotations in T2T ape genomes (in Mb and percentage of total genome length), shown separately for autosomes and sex chromosomes. APR: A-phased repeats; DR: direct repeats; STR: short tandem repeats; IR: inverted repeats; MR: mirror repeats; G4: G-quadruplexes; Z: Z-DNA. Note that the scale on the X-axis is different for each chromosome type. The data for this figure are in Table S1A. The statistics for the triplex motif is in Table S1B.

Overlapping annotations of different non-B DNA motif types. We found substantial overlap among non-B DNA annotations of different types (Fig. 3, Fig. S2), suggesting alternative structure formation afforded by the same genomic sequence. For instance, ~70%, ~55%, and ~48% of Z-DNA annotations on the human autosomes overlapped with STR, mirror repeats, and direct repeat annotations, respectively (Fig. 3B). The amount and types of motifs that overlapped differed between autosomes and sex chromosomes. For autosomes, the largest overlap was found between direct repeats and STRs, followed by the overlap between these two types and mirror repeats. For the Y chromosome, the largest overlap was found between mirror and inverted repeats, with overlapping annotations spanning more bases than non-overlapping annotations for these motifs (Fig 3A). Non-human apes showed patterns of overlap similar to those observed in humans for both the autosomes and the X chromosome. However, differences were observed for the Y chromosome (Fig. S2A-L). For example, the overlap between mirror and inverted repeats on chromosome Y was less pronounced in non-human apes than in humans.

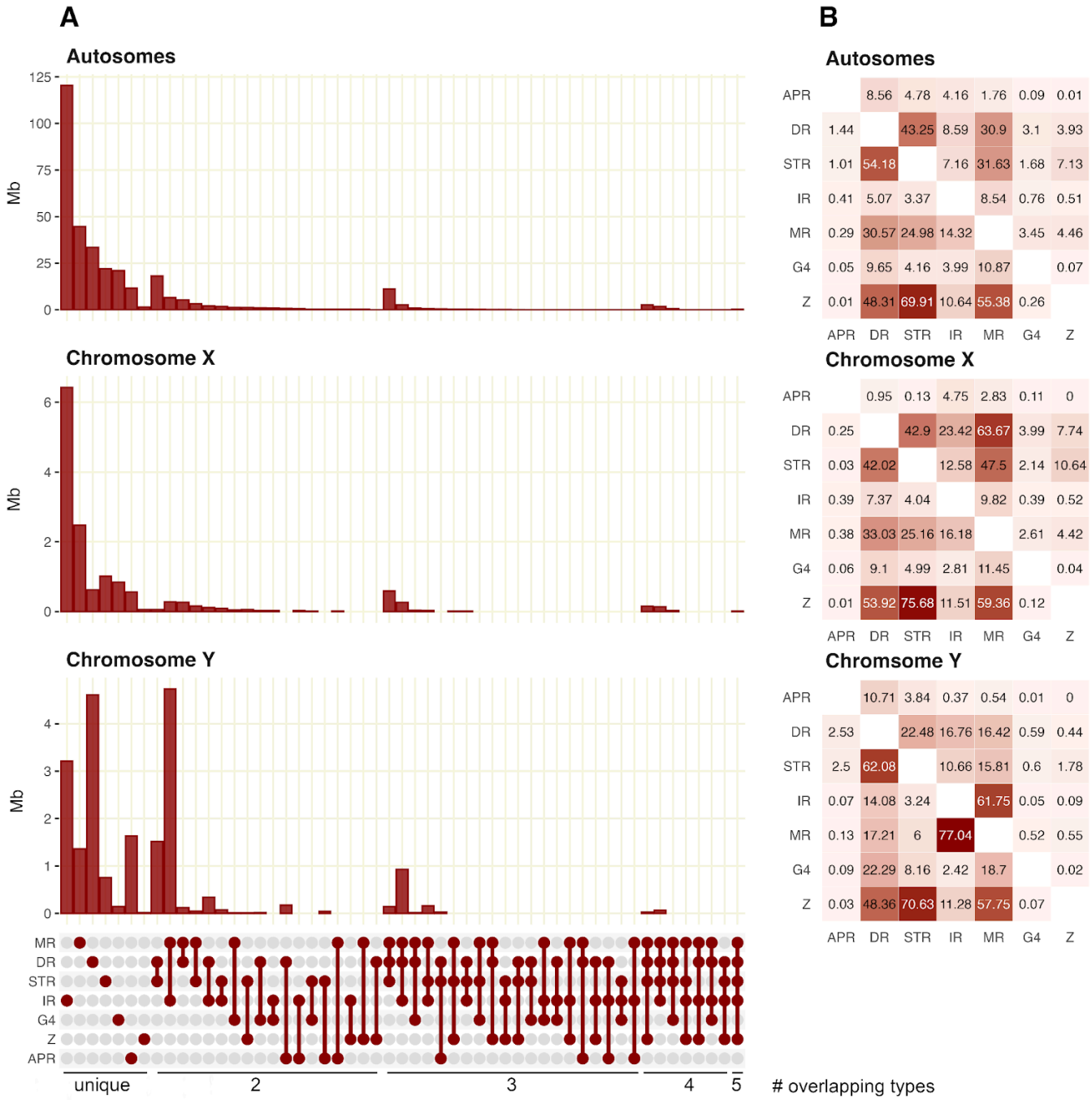


Figure 3. Non-B DNA motif types annotations and their overlaps (i.e., the same bases annotated) in human T2T autosomes, chromosome X, and chromosome Y. (**A**) Number of Megabases and upset plot, comprising all combinations with a total overlap >10 kb, (**B**) Pairwise overlap given as

the percentage of the row type (indicated on the left) that overlaps with the column type (indicated at the bottom). APR: A-phased repeats; DR: direct repeats; G4: G-quadruplexes; IR: inverted repeats; MR: mirror repeats; STR: short tandem repeats; Z: Z-DNA. See Fig. S2 for non-B DNA motif types annotations and overlaps in the other species.

Distribution of non-B DNA motifs along the chromosomes: General trends. A visual inspection of the density of non-B DNA motifs along ape chromosomes (Fig. 4, Fig. S3, and Fig. S4) suggested the following trends. In humans, all non-B DNA motif types have high density on the short arms of acrocentric chromosomes (chromosomes 13, 14, 15, 21, and 22), and A-phased, direct, short tandem, inverted, and mirror repeats have high density in the heterochromatic region of the Y chromosome (Fig. 4C). The acrocentric chromosomes in non-human great apes showed similarly high densities of non-B DNA motifs on the short arms, especially for direct repeats and STRs in gorilla and orangutans (Fig. 4D-F). The patchwork of different non-B motifs corresponded very well to the centromeric satellite repeat annotation, with, for example, HSAT1 enriched in inverted and mirror repeats and HSAT3 enriched in A-phased, direct, and short tandem repeats (Fig. S5, Fig. S6). rDNA was enriched in all non-B types except A-phased and inverted repeats. Subtelomeric regions were frequently enriched in G4s. An interesting pattern was observed at and around the centromeres, with some centromeres showing high densities for certain non-B DNA motifs, while others having higher densities in the flanking regions. We investigated some of these patterns in more detail below.

Enrichment of non-B DNA motifs at genes and regulatory elements. To perform a more rigorous analysis of non-B DNA enrichment, we evaluated it in different functional regions, repeats (based on RepeatMasker annotations), and the remaining, presumably non-functional non-repetitive regions of the human genome (similar analyses were not performed in non-human ape genomes due to incomplete annotations of functional sequences). Many types of non-B DNA were previously 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 enrichment of G4s and Z-DNA at promoters and enhancers, as well as at origins of replication (Fig. 5). CpG islands were enriched in all types of non-B DNA motifs except for A-phased and inverted repeats. 5' untranslated regions (UTRs) and, to a smaller degree 3'UTRs, as well as coding sequences, were enriched in G4s. This was still true after correcting G4 fold enrichment for GC content using a simple conversion factor (see Methods, Fig. S7, and Discussion).

Enrichment of non-B DNA motifs at repeats and satellites. The T2T genomes provided a complete resolution of repeats in the ape genomes, including transposable elements and satellites, allowing us to evaluate non-B DNA present at such genomic regions comprehensively. In general, repetitive sequences harbored more non-B DNA than non-repetitive sequences (for example 1.4× more in human, and 2.0× more in gorilla,Table 2). Simple repeats and low-complexity regions were strongly enriched in most types of non-B DNA motifs. Considered together, transposable elements were not enriched for non-B DNA (Fig. 6,Table S2). However, some transposable elements showed enrichment in inverted and A-phased repeats, and SVA was enriched in G4s, and direct and mirror repeats. Interestingly, RNA as a group were enriched in G4s (Fig. 6B), a signal driven by ribosomal RNA (rRNA, Fig. 6A).

As a group, satellites were enriched in A-phased, direct, inverted, and mirror repeats, as well as in STRs, but not in G4s and Z-DNA, for all great apes studied but not for the siamang (Fig. 6B, Table S2). The patterns of non-B DNA enrichment at particular satellites (see Fig. 6A and Fig. S8 for human and Fig. S9 for non-human apes) were often shared across species. For instance, the LSAU and MSR1 satellites were enriched in G4s in all studied species. The Nereid, Neso, and Proteus satellites were enriched in direct, inverted, and mirror repeats in all studied non-human apes (Fig. S9A-F). The human repeat annotations have been recently updated with manually curated satellites and composite repeats[67]. Many of these satellites were enriched for direct repeats. However, surprisingly, we found no enrichment of G4 motifs in Walusat (Fig. S8A), which was reported previously[67]. Composite repeats were more often enriched in G4 and Z-DNA motifs than in the other non-B DNA types (Fig. S8B).

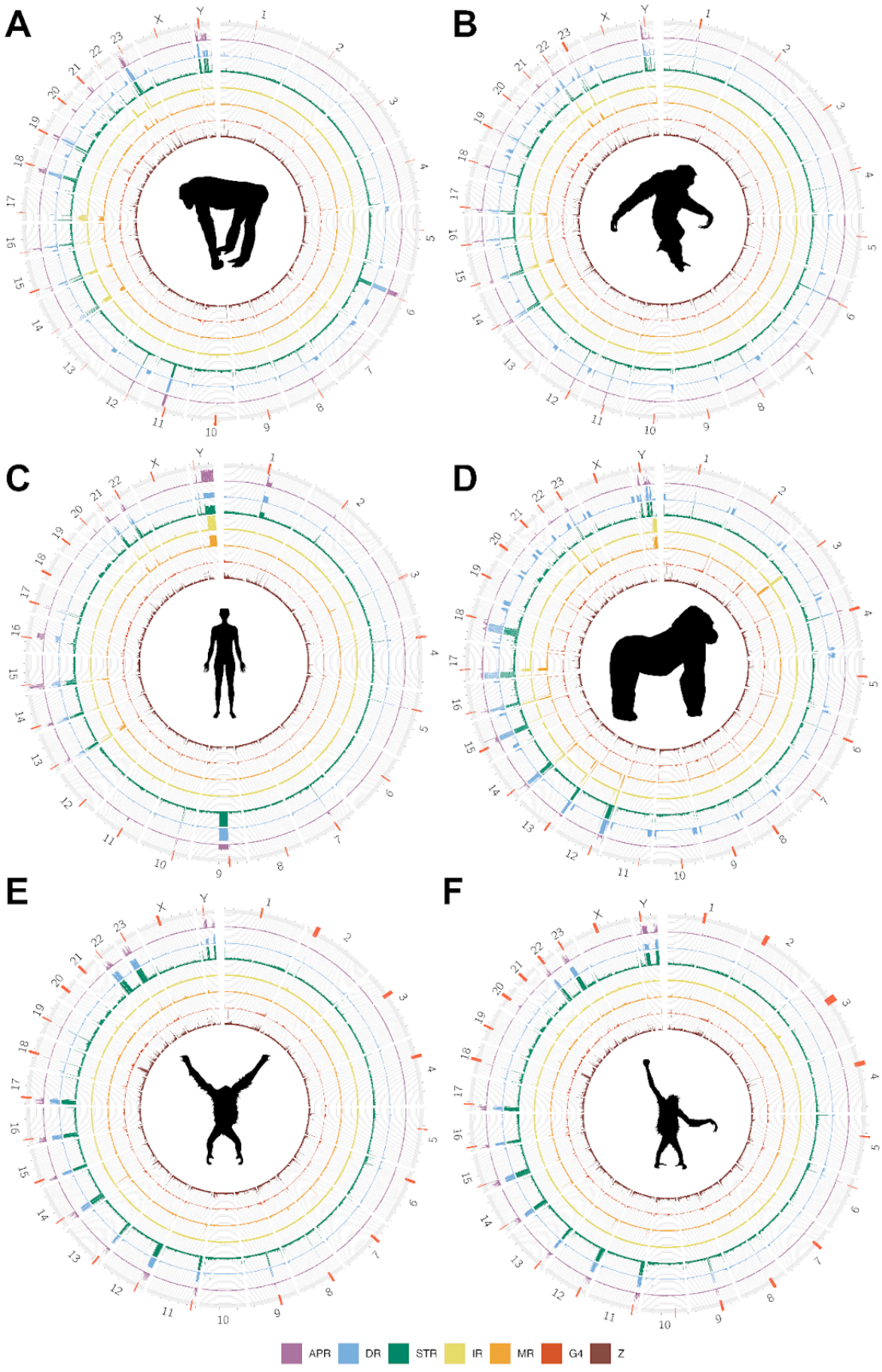


Figure 4. Non-B DNA density along the primary haplotype chromosomes for (A) chimpanzee, (B) bonobo, (C) human, (D) gorilla, (E) Bornean orangutan, and (F) Sumatran orangutan. See Fig. S3 for non-B DNA density in siamang. Active centromeres are marked with red stripes along the chromosomes. Note that an active centromere was not found for chr10 in Bornean orangutan. Abbreviations for non-B DNA are as in Fig. 2. The alternative haplotypes are shown in Fig. S4. Animal silhouettes are from https://www.phylopic.org.

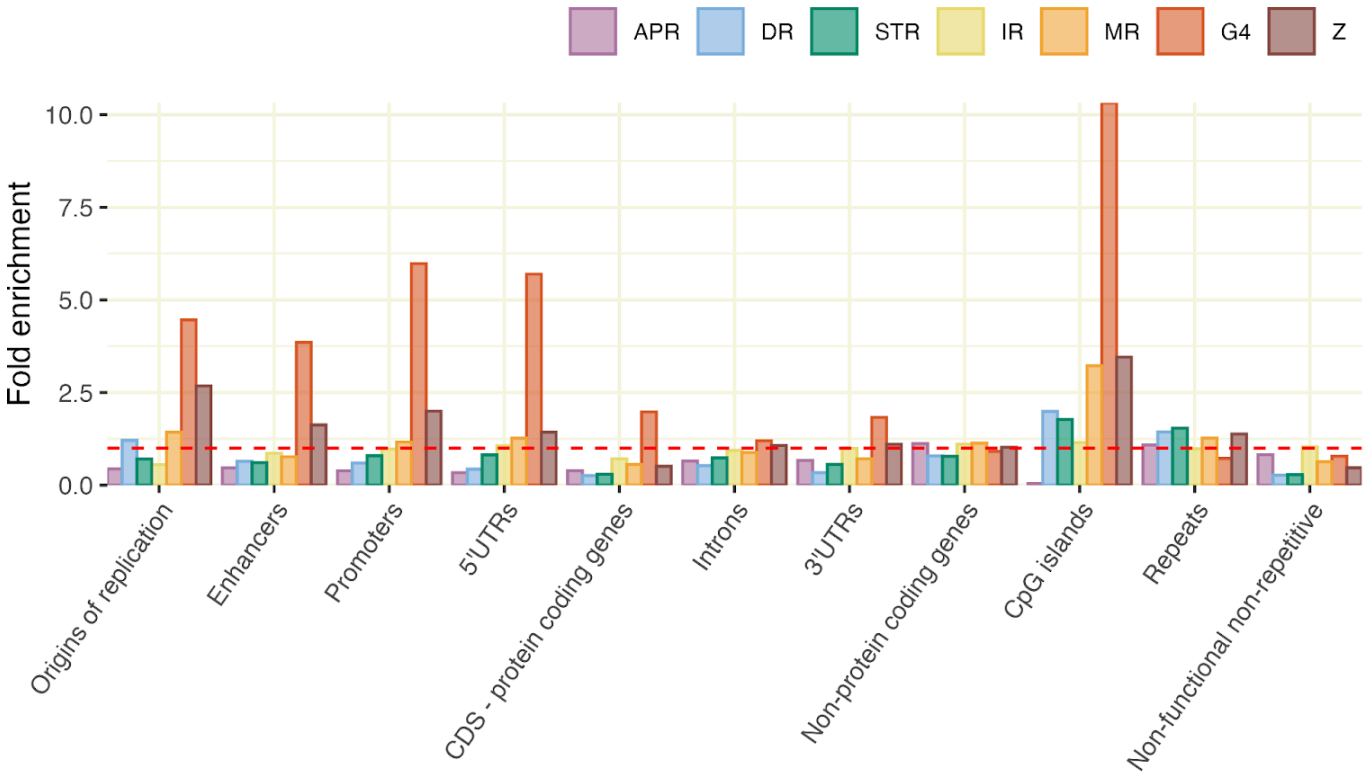


Figure 5. Enrichment of non-B DNA motifs at different functional regions of the genome. Fold enrichment is calculated compared to genome-wide density. Non-functional non-repetitive regions represent sequences that do not belong to the other categories. Red dashed line (fold enrichment=1) represents the genome-wide average. Abbreviations for non-B DNA are as in Fig. 1.

Table 2. Enrichment of non-B DNA motifs at repetitive sequences. Fold enrichment is calculated as non-B motif density at repeats compared to density in non-repetitive sequence, using only the primary haplotypes. Abbreviations for non-B DNA are as in Table 1.

Species	APR	DR	STR	IR	MR	G4	Z	all
Bonobo	1.31	5.90	5.52	1.04	1.90	0.55	2.15	1.53
Chimpanzee	1.08	4.73	4.10	0.97	1.88	0.55	2.22	1.36
Human	1.58	4.47	5.15	1.00	1.92	0.56	2.28	1.39
Gorilla	0.91	11.32	7.86	1.05	2.12	0.66	1.92	2.04
Bornean orangutan	1.10	4.93	4.65	0.86	1.76	0.58	2.15	1.33
Sumatran orangutan	1.13	4.56	4.47	0.89	1.76	0.57	2.13	1.31
Siamang	0.78	2.61	3.37	0.92	1.49	0.51	1.94	1.03

We further analyzed some human satellites and repeats enriched in G4 motifs (Table S3). These included retrotransposon SVA, rRNA, and satellites ACRO1, GSAT, GSATII, GSATX, HSAT5, LSAU, MSR1, SAT-VAR, SST1, and TAR1. Whereas G4 motifs were enriched in these cases, we had no evidence of G4 formation vs. non-formation. As a proxy of such formation, we evaluated methylation profiles for G4s and repeats/satellites harboring them in the HG002 lymphoblastoid and the CHM13 hydatidiform mole cell lines[68], as methylation was shown to be antagonistic to G4 formation[69,70]. In the SVA retrotransposon, both cell lines were methylated at G4s as well as at the full repeat region, suggesting that G4 structures do not form at this retrotransposon. In most of the other cases we considered, G4 motifs enriched at repeats and satellites were unmethylated in CHM13 and methylated in HG002, reflecting the overall methylation trends at these genomic regions for these cell lines[68] (Fig. 6C). However, for certain satellites (e.g., LSAU and TAR1 in HG002, and LSAU and SST1 in

CHM13), the methylation level for G4s was lower than that for the satellites harboring them, suggesting G4 structure formation. For some satellites (e.g., HSAT5 in HG002 and MSR1 for both cell types), we observed a bimodal distribution of methylated vs. unmethylated G4s, suggesting that alternative structure formation (i.e., B vs. non-B DNA). Similarly, rRNA, and G4 motifs in it, were largely unmethylated in CHM13, suggesting G4 formation, and had a bimodal methylation score distribution in HG002.

We then experimentally validated three G4 motifs found in the LSAU satellite—one of the satellites with the strongest G4 enrichment in our dataset—and the previously reported potentially G4-forming sequence in Walusat[67] (not annotated by Quadron, see Methods). Circular dichroism spectroscopy (CD), isothermal difference spectra, and thermal difference spectra consistently supported the formation of intramolecular parallel-stranded G4s for two of the three LSAU motifs tested. In contrast, the third LSAU motif formed a hairpin structure. The sequence from Walusat formed canonical B DNA in all assays (Fig. 6D, Fig. S10).

We also investigated the satellite SST1 in more detail, since it was recently suggested to act as the breakpoint in Robertsonian translocations of acrocentric chromosomes 13, 14, and 21 in humans[71,72]. The satellite itself was enriched in G4s (~3.4-fold higher density than the genome-wide average), with some differences between the satellite subtypes (Table S4). In particular, subtypes sf1, present on acrocentric p-arms, and sf2, present on chromosomes 4, 17, and 19, were enriched in G4 motifs, whereas subtype sf3, present on chromosome Y, lacked such an enrichment. Strikingly, the sequence between the monomers was found to be enriched in non-B motifs, especially in Z-DNA, for all SF subtypes. Notably, this enrichment was most prominent at subtype sf1 (spacer length of ~135 bp), where it was 97×, 15×, 7×, and 5× for Z-DNA, mirror repeats, direct repeats, and STRs, respectively, compared to the genome-wide average (Fig. S11, Table S4). The presence of non-B DNA motifs at such high density could be involved in the destabilisation of these regions (see Discussion).

Enrichment of non-B DNA motifs at centromeres. We performed a more detailed analysis of the enrichment of non-B DNA at experimentally annotated, active centromeres available for the six great apes in our data set (Fig. 7, Fig. S12, and Fig. S13), as it was suggested that non-B DNA determines centromere formation[35]. Overall, most centromeres (158/263, or 60%) showed significant enrichment in at least one type of non-B DNA, and over a quarter of centromeres (74/263, or 28%) had >2-fold enrichment (Fig. 7, Fig. S12). However, G4s were always underrepresented at centromeres. The other types of non-B DNA displayed species- and chromosome-specific trends. Inverted repeats displayed a moderate enrichment at approximately half of all centromeres. A-phased repeats showed significant enrichment at centromeres of some chromosomes in most species, with a particularly high enrichment and many chromosomes affected in chimpanzee and human. Z-DNA showed enrichment at centromeres of some chromosomes in bonobo, chimpanzee, and human. In contrast, mirror repeats showed enrichment at centromeres of some chromosomes in the two orangutan species.

Within species, some of these patterns could be explained by suprachromosomal families (SFs)[61,73]. Taken all species together, SF01 centromeres were enriched in Z-DNA, SF1 centromeres were usually enriched in inverted repeats, and SF4—in A-phased repeats (Fig. S14). Some patterns were species-specific though. The centromeres in the two orangutan species almost exclusively belonged to SF5 and slightly more than half of them exhibited enrichment in non-B DNA (SF1-3 dates after orangutan split from the other great apes[73]). In gorilla, chromosomes with centromeres annotated as SF1 were enriched in inverted repeats, while in chimpanzee and bonobo, SF1 chromosomes showed a mixed pattern of either Z-DNA enrichment, or A-phased, inverted, and/or mirror repeat enrichment (Table S5). The human SF1 and SF2 chromosomes generally lacked non-B DNA enrichment, with a notable exception of chromosomes 13 and 21, which were classified as SF2 and were enriched in A-phased repeats.

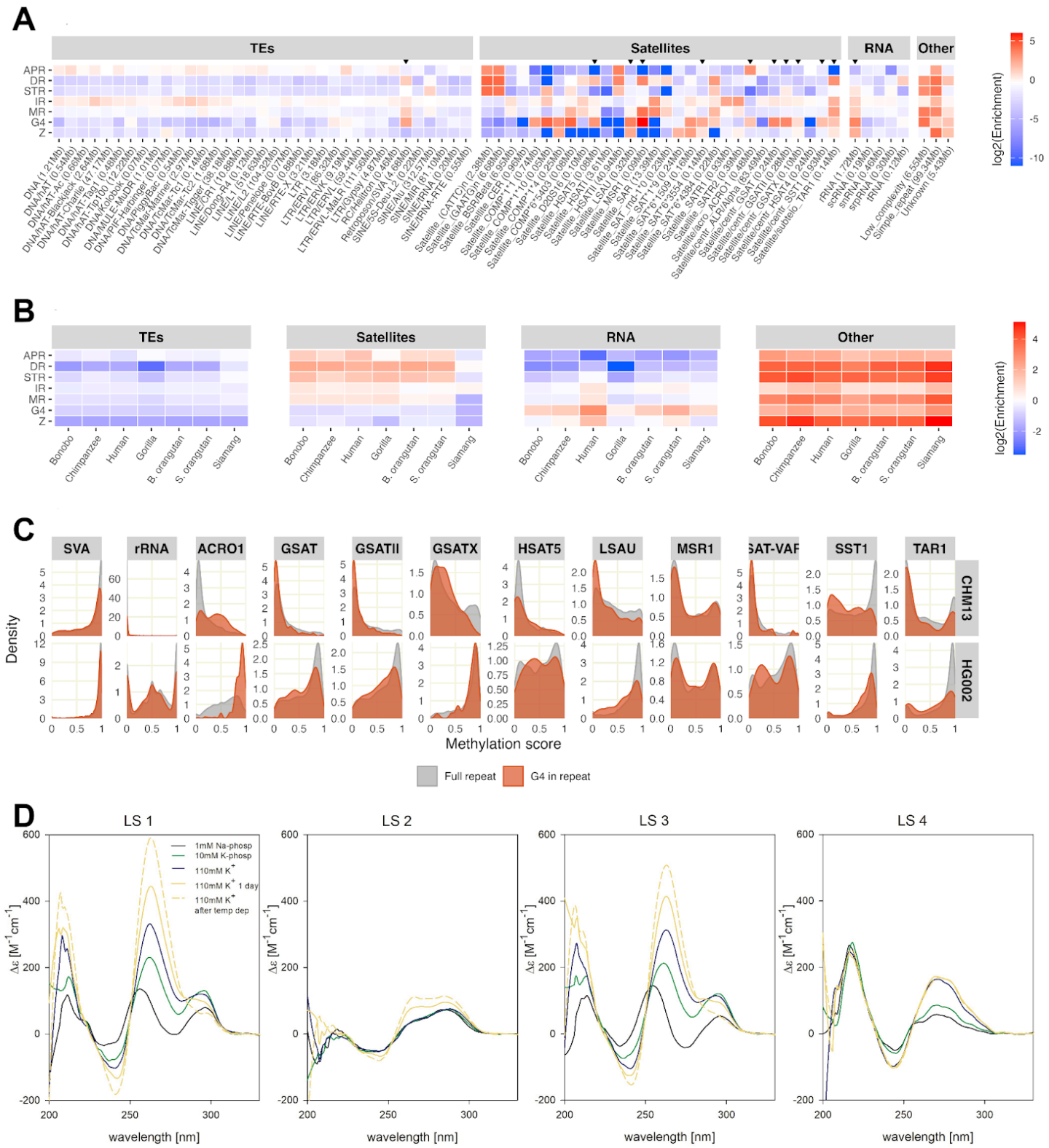


Figure 6. Non-B DNA at repeats and satellites. (**A**) Enrichment of non-B DNA motifs at repeats, given as log-fold densities compared to genome-wide densities for the human genome. Underrepresentation of non-B DNA (values below 1) is shown in blue, whereas enrichment (values above 1) is shown in red. Long repeat names were 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. Abbreviations for non-B DNA are as in Fig. 2. Repeats marked with a black arrow are further investigated in part C. (**B**) Fold enrichment (compared to genome wide) in the four repeat groups for all seven species, centromeric satellite repeats (shown in Fig. S6) are included in the Satellite group. (**C**) Distribution of methylation scores at selected repeats (gray) and G-quadruplexes (vermilion) overlapping with the repeat annotations, obtained from CpG sites in the human lymphoblastoid cell line (HG002) and hydatidiform mole cell line (CHM13). A score of 1 means that all reads were methylated at this site, whereas a score of 0 means no methylation. (**D**) Circular dichroism spectroscopy of the four experimentally tested G4 motifs: LS1-LS3 from LSAU and LS4 from Walusat. Samples were measured in the potassium chloride buffer (110 mM K^+= 10mM K-phosphate, pH 7 + 95 mM KCl) in 1μM strand concentration.

When comparing the non-B DNA patterns in centromere regions with detailed centromeric satellite repeat annotation from[61,73], we observed that the regions of high non-B density often overlapped perfectly with the satellite annotations (Fig. S13), similar to what was observed in the short arms of acrocentric chromosomes. For example, A-phased, direct, and short tandem repeats often occurred at HSAT2 and HSAT3 satellites, while HSAT1A mostly overlapped with inverted and mirror repeats. Inactive α-satellites (a higher order repeat, or HOR, that does not interact with the kinetochore[73]), divergent α-satellites (older HORs that have started to degrade), and monomeric α-satellites (not organised into HORs) were in general not enriched for any type of non-B DNA motifs.

In all species, active centromeres annotated in the primary and alternative haplotypes showed similar non-B DNA enrichment and depletion patterns, with some striking exceptions. In bonobo, the centromere on the primary haplotype of chromosome 17 was significantly enriched for Z-DNA, while such enrichment was missing entirely in the centromere of the alternative haplotype of chromosome 17. We note that the active centromeres on these two haplotypes belong to different suprachromosomal families (SF4 and SF1 for the primary and alternative, respectively). The same pattern of enrichment and depletion of Z-DNA was observed between the centromeres on primary and alternative haplotypes of chromosome 15 in chimpanzee, however this discrepancy cannot be explained by different SF families, as both haplotypes were annotated as belonging to SF1.

Bornean orangutan chromosome 10 lacked an annotated active centromere for the primary haplotype (no evidence of CENP-A enrichment on the alpha satellite HOR array[61]). As there was an annotated centromere for the alternative haplotype, we sought to compare the non-B density in this region between the two haplotypes. However, the alignment of the region between the two haplotypes revealed that the centromere from the alternative haplotype, as well as the 1-Mb upstream flanking region, were entirely missing from the primary haplotype (Fig. S15). The alternative haplotype centromere was depleted of non-B DNA compared to the genome average, but the 1-Mb upstream flanking region showed enrichment in several types of non-B DNA motifs, especially in A-phased repeats, direct repeats, and STRs.

We also sought to compare the non-B density in centromeres across the ape species with and without CENP-B binding motif (‘CENP-B box’), since it has been shown that the lack of CENP-B binding is correlated with increased non-B DNA formation[35]. We found little difference in non-B DNA motif content between the 249 centromeres containing the CENP-B box motif vs. the 23 centromeres that lack the motif (Fig. S16B). However, the 1-Mb flank on the p-arm showed significantly higher enrichment for A-phased repeats, direct repeats, and short tandem repeats in centromeres without the CENP-B box (Fig. S16B). Additionally, the densities of inverted repeats, mirror repeats, and Z-DNA were significantly lower in the 1-Mb flanks of centromeres without vs. with the CENP-B box. Note that for both groups of centromeres, the flanks were depleted in these non-B DNA motifs compared to their average frequency genome-wide. The q-arm flanks showed no significant differences between the two groups (Fig. S16C).

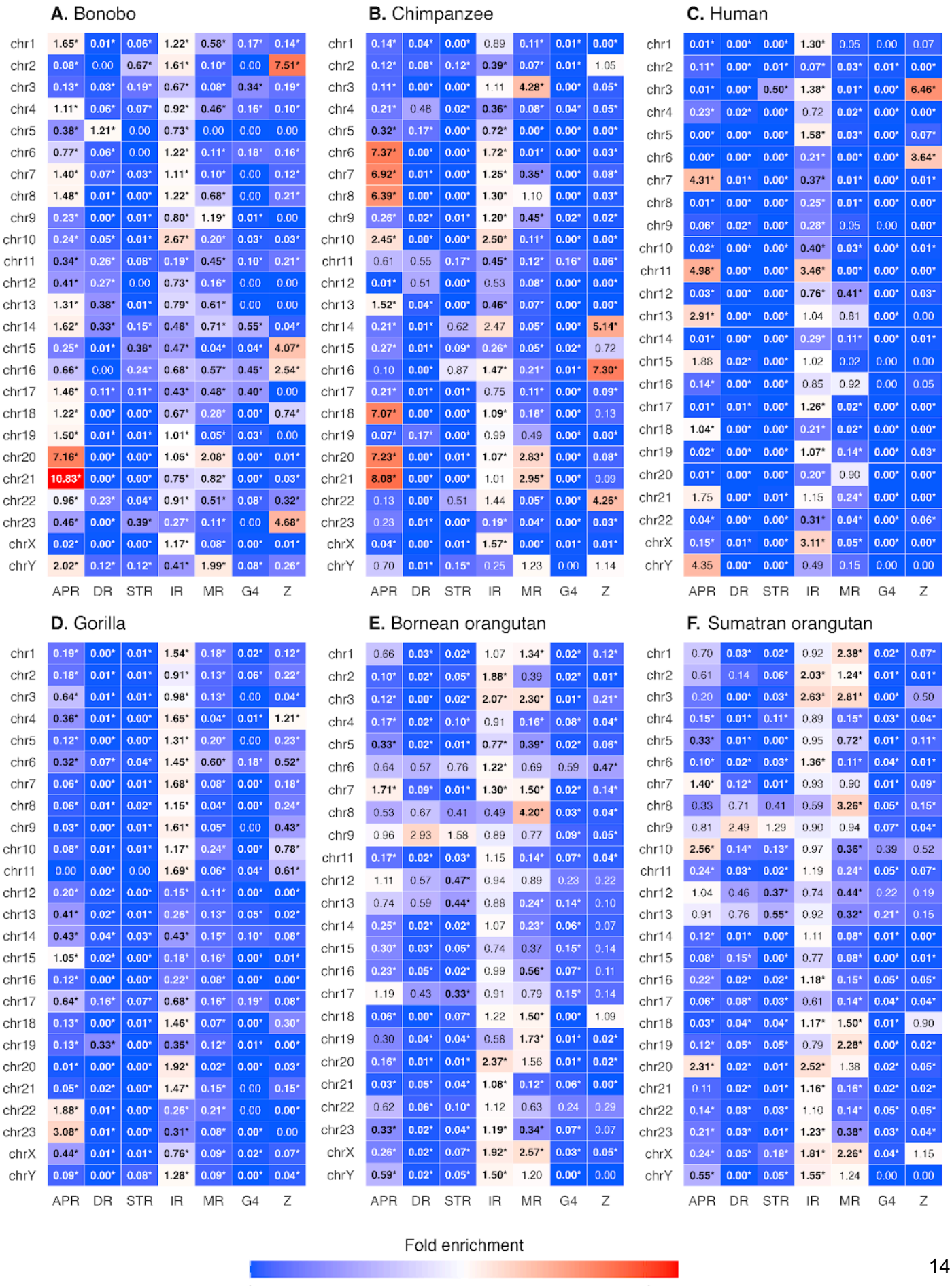


Figure 7. Fold-enrichment of non-B DNA densities in primary haplotype centromeres as compared to genome-wide average densities for (A) bonobo, (B) chimpanzee, (C) human, (D) gorilla, (E) Bornean orangutan, and (F) Sumatran orangutan. The underrepresentation of non-B DNA (values below 1) is shown in blue, while enrichment (values above 1) is shown in red. Densities with a significant underrepresentation or enrichment compared to the genome-wide average density are marked in bold with '*' (two-sided randomization test, $P<0.05$). Abbreviations for non-B DNA are as in Fig. 2. Fold-enrichment for alternative haplotype centromeres can be found in Fig. S12.

Discussion

We conducted a detailed analysis of non-B DNA motifs in the T2T assemblies of human and non-human ape genomes, which have recently become available[58–61]. Importantly, these genomes have been produced with the use of long-read sequencing technologies, which are known to be less error-prone at non-B DNA motifs compared to the Illumina short-read technology [56]. Additionally, due to the use of long reads and novel assembly algorithms, these genomes have resolved highly repetitive genomic regions, such as long satellite arrays, including complete centromeres. We 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[60] and autosomes[61]. The ability to analyze previously inaccessible regions of the genomes, which are rich in non-B DNA motifs, allowed us to uncover the complete genome-wide repertoire of such motifs in humans, as well as non-human apes whose T2T genomes are available to date. Thus, our study complements earlier 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 are unevenly distributed among and along ape genomes. In the human genome, many blocks of high-density non-B DNA motifs correspond to centromeric satellites or satellites at the short arms of acrocentric chromosomes. We found that short arms of acrocentric chromosomes correspond to a patchwork of different combinations of non-B DNA motif types. For instance, the HSAT1 satellite is rich in inverted and mirror repeats, and the HSAT3 satellite is rich in A-phased repeats, direct repeats and 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 inter-unit similarity. Therefore, these non-B DNA features should be incorporated into future models of satellite evolution[75]. When considered as groups, satellites displayed enrichment in most non-B DNA motif types, whereas transposable elements did not show such an enrichment.

We identified several instances of non-B DNA motif enrichment at particular satellites and transposable elements, consistent with previous analyses of non-T2T genomes showing that some types of non-B DNA motifs are present at repeats, and might be propagated through their spreading[51,76–78]. For G4 motifs in particular, we were able to predict formation based on methylation status (methylation inhibits G4 formation[70]). In many instances, G4s enriched at transposable elements (e.g., at SVAs) and satellites were methylated, and thus unlikely to form. Note that, in contrast to this pattern, G4s were shown to form at the SVA inserted in the *TAF1* gene and affect its expression in patients with X-linked dystonia parkinsonism[79]. However, in some instances, G4s were less methylated than the overall satellites they are embedded into, or had a bimodal methylation density distribution. 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.

LSAU was one of the satellites with high G4 enrichment in our dataset, and we validated G4 formation in it experimentally. This satellite has previously been shown to have variable methylation levels in apes[80] and speculated to have an effect on gene expression[81]. It is also part of the larger repeat complex D4Z4 whose copy number and methylation level are associated with fascioscapulohumeral muscular dystrophy[82]. Our *in vitro* experiments confirmed the formation of two out of three tested LSAU G4 motifs showing low methylation in the CHM13 cell line. We note that, compared to the other two motifs tested, the LSAU motif that did not form a G4 *in vitro* showed higher methylation in HG002 and had lower Quadron stability score (19.31, compared to >31 for the other two, and very close to the threshold of 19 suggested by the Quadron authors for discriminating stable and unstable G4s[65]) than the two others motifs tested. This motif also contained many cytosines that the guanines could pair with in a hairpin, rather than forming a G4 (see Methods).

We found no enrichment of G4 motifs in Walusat, a satellite that previously has been reported as enriched in this motif[67]. This discrepancy might result from the use of different prediction software programs. Quadron only predicts standard motifs with four G3 stems, while Hoyt and colleagues[67] based their predictions on G4Hunter[83], which additionally includes G4 motifs with bulges (i.e., the G3 stem can be interrupted by other nucleotides); this is the G4 type found in Walusat. We repeated the G4 prediction of the Walusat array on chr14 from[67] and found that the G4Hunter scores of the four most common G4 motifs (each occurring >2,500-5,000 times at Walusat occurrences across the genome) were low, i.e. in the range of 1.20-1.32. This is very close to the default threshold (1.2) and below the more stringent threshold of 1.5 suggested to reduce the false discovery rate to below 10%[84]. Our experimental validation of the most common Walusat motif resulted in B DNA formation. We conclude that G4 structures are unlikely to form at Walusat.

We performed an in-depth investigation of the satellite SST1, which is present in large arrays on the short arms of acrocentric chromosomes and was suggested to be the breakpoint of Robertsonian translocations in humans[71,72]. We discovered that non-B motifs are enriched not only at the annotated SST1 satellites themselves, but also at the sequence between its satellite monomers. The SST1 subtype sf1, which is present on the p-arms of chromosomes 13, 14 and 21, has a binding site for PRMD9, a recombinogenic protein. It was suggested that the resulting increase in recombination is one of the prerequisites for the Robertsonian chromosome formation[72]. Here, we show that the spacers between the SST1 satellite monomers are highly enriched in Z-DNA, which is another known inducer for double-strand breaks[85], and that this enrichment is by far highest on the aforementioned acrocentric chromosomes. We hypothesize that this enrichment could also play an important role for this type of translocation.

Our examination of active, experimentally defined centromeres in great apes indicated that more than half of them are enriched in at least one type of non-B DNA motifs, particularly A-phased and direct repeats. This extends an earlier study of non-B DNA enrichment at active centromeres of human, African monkey, and mouse[35] to complete chromosome sequences of multiple species of great apes and suggests an important role of non-B DNA structures in defining centromeres. In fact, Patchigolla and Mellone[38] studied fruit fly chromosomes and suggested that satellite repeats occur at centromeres at least in part because they can form non-B structures. Enrichment in non-B DNA motifs and in R-loop formation was also found at oat centromeres[39], arguing that the involvement of non-B DNA in centromere definition and/or function might be conserved across eukaryotes.

Whereas we observed a pattern of non-B motif enrichment at the centromeres, we could not clearly detect a particular non-B DNA type being the dominant feature of centromeres. Instead, many centromeres were annotated as harboring several non-B DNA types. This is consistent with a recent analysis of human centromeres suggesting that alternative non-B structures can form at them, as evident from high ensemble diversity values[86]. Centromeres belonging to the same SF often (but not always) shared common patterns of non-B DNA enrichment. We note that the annotation of SF into subtypes was developed for the human genome, and hypothesize that a more detailed annotation of the non-human apes will generate more subtypes of suprachromosomal families and further increase the correlation between non-B DNA and SFs.

We saw no difference in non-B DNA enrichment between centromeres containing CENP-B binding motifs compared to centromeres lacking these motifs, while the p-arm 1-Mb flanks of centromeres without CENP-B showed enrichment for several non-B motif types. The CENP-B binding motif is highly conserved over many taxa and has been shown to be essential for *de novo* centromere formation on synthetic chromosomes[87]. However, it is absent from some centromeres, and Kasinathan and Henikoff[35] suggested that non-B DNA can substitute CENP-B motif in them. Perhaps, to define the centromere, non-B DNA does not have to form within the active α-satellite itself, but can instead occur in its close proximity. In fact, a recent study investigating the minimum free energy (MFE) and thermodynamic ensemble diversity as a proxy for secondary structure and stability in human centromeres found highest MFE (indicating low stability and non-canonical structure formation) in both the active centromere itself and the divergent HORs adjacent to it[86]. This is consistent with the importance of the pericentromeric regions for keeping the sister chromatids together at meiosis[88], a process possibly mediated by alternative DNA structures.

We showed that several functional elements, including enhancers and promoters, are enriched in G4s, consistent with findings in a previous, non-T2T version of the human genome[74] and in other taxa (reviewed in

[89]). These regions are often GC-rich, and GC content correlates with G4 motif abundance (see for example [90]). Nevertheless, it is not resolved whether G4 motifs are often found in these regions because they are GC-rich, or whether these regions are GC-rich due to their high G4 density. We showed that the G4 enrichment for most functional elements also remained after a linear GC correction we applied as in[74]. However, as the GC-G4 correlation might be non-linear [91], a more sophisticated correction might be required. For instance, Mohanty and colleagues[91] found that coding regions are no longer significantly enriched in G4s after a quadratic correction for GC content. Because the other non-B motif types show relationships with GC content that are neither linear nor quadratic, we had difficulty in finding the same GC correction model suitable for all seven non-B motif types investigated in this study.

One caveat of our study is that the software we used, gfa, only predicts non-B DNA motifs with identical arms for direct, inverted, and mirror repeats. On the one hand, this explains why not all satellites are annotated as direct repeats, even though most have monomers shorter than the maximum arm length considered by gfa (300 bp). On the other hand, mismatches in the arms sequence should destabilize the potential formation of non-B DNA. Different types of non-B DNA motif annotations for the same sequences, however, can lead to more stable non-canonical structures. For instance, slipped-strand structures in long sequences of STRs, which in turn contain inverted motifs (e.g. $CTGCAG_n$), are known to be stabilized by the formation of hairpins in the loops[92]. Here we included all mirror repeats in our analyses, to be consistent with several prior studies[2,60,86,93,94]. We note that only a subset of mirror repeats is predicted to form triplex DNA (Table S1B).

In the future, 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 structures these motifs form, 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 the human *c-MYC* oncogene can form either a G4 or H-DNA, which might have different effects on genomic instability in this genomic region[22]. Similarly, knowing what particular non-B DNA structures form at satellites can inform their expansion mode.

In conclusion, our new annotations of non-B DNA motifs in complete ape genomes have shown that there is strong but uneven potential for non-B formation along these genomes and among species. This potential was particularly high in the genomic sequences added to the T2T assemblies. We predict formation of several alternative secondary structures at many genomic locations. Further studies and experimental validation will determine which of these structures form in any given species and tissue, as well as their effects on cellular processes.

Methods

Non-B DNA annotation. Non-B DNA motifs were annotated for bonobo (*Pan paniscus*), chimpanzee (*Pan troglodytes*), human (*Homo sapiens*), gorilla (*Gorilla gorilla)*, Bornean orangutan (*Pongo pygmeaus*), Sumatran orangutan (*Pongo abelii*), and siamang (*Symphalangus syndactylus*), as described in[61]. In short, motifs of A-phased repeats, direct repeats, mirror repeats, STRs, and Z-DNA were annotated in each T2T genome with the software gfa[7] (https://github.com/abcsFrederick/non-B_gfa) with the flag *–skipGQ*. This predicts A-phased repeats with at least three A-tracts of length 3-9 bp and 10-11 bp between A-tract centers (it also looks for T-tracts that correspond to APRs on the reverse strand); direct repeats with lengths 10-300 bp and a maximum spacer length of 100 bp (we note that no DR spacer was longer than 10 bp, Fig. S1); inverted repeats with arms of 6 bp or longer and a maximum loop size of 100 bp; mirror repeats with arms of 10 bp or longer and a maximum loop size of 100 bp; STRs with repeated units of size 1-9 bp and a total length of at least 8 bp, and Z-DNA motifs (alternating purine-pyrimidine nucleotides longer than 10 bp). Triplex motifs were extracted from the mirror repeats ('grep subset=1', default parameters of minimum purine/pyrimidine content of 10% and maximum spacer length 8 bp were used to define the subset). G4s were annotated using Quadron[65] with default settings, which predicts standard G4s with at least four GGG-stems without bulges. The output from each motif type was converted to bedformat, and any overlapping annotations of non-B DNA motifs of the same type were merged with mergeBed from bedtools v 2.31.1[95]. For G4s, 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 [61]) for which there was an older non-T2T genome available, was 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[96]. We followed the winnowmap recommendations and first generated a set of high-frequency *k*-mers with meryl v1.4.1 [97] 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 not be considered new). Densities of non-B motifs in 'new' and 'old' sequences (i.e., sequences in T2T genomes that did not align vs. aligned to the older assembly versions, respectively) 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 sequences were compared with a chi-square goodness of fit test for each non-B motif type and chromosome type separately, and Bonferroni-corrected for multiple testing. To assess results robustness, we also randomly resampled half of the data 10 times and repeated the chi-square goodness of fit tests.

Enrichment in functional regions. Gene annotations for human (CHM13v2.0) were taken from[91]. We considered G4s annotated on both strands. For the other non-B DNA types, the annotations are the same for both strands. Fold enrichment was calculated as non-B motif density for each region divided by the genome-wide non-B DNA density. Since G4s are more likely to form in GC-rich regions, we corrected the enrichment in this motif category by multiplying it by a correction factor, following an approach used in[74]

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[67] were downloaded from the same source and analyzed separately. RepeatMasker output was converted to bed format 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 was compared to the genome-wide density using python and bash scripts (provided on the github). Methylation data for 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, and methylation data for the CHM13 cell line was downloaded from chm13v2.0_CHM13_CpG_ont_guppy3.6.0_nanopolish0.13.2.bw[68]. These files contain methylation scores for CpG sites, given as a fraction of methylated reads (0 means no methylation was detected in any reads, and 1 means all reads were methylated). We extracted G4s 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 G4 motifs with the distribution of methylation scores from all annotated repeats of each repeat class. To investigate the discrepancy in predicted G4s for the Walusat repeat in humans between our study and [67], we repeated their G4Hunter analysis (through the web application[83], https://bioinformatics.ibp.cz/#/analyse/quadruplex) on chr14:260778-634253 using default settings and downloaded the resulting G4 hits as a csv file. As G4Hunter only reports results for 25-nt windows, we parsed the Walusat fasta sequence and split it at a commonly occurring pattern (GGGGTCA, chosen so that the sequences start with the longest stretch of guanines). This resulted in the majority of sequences of 64 nt (the Walusat repeat length), and a minority of sequences longer than 64 nt (for diverged monomer copies lacking the aforementioned pattern). We cut all sequences at 64 nt, sorted them, and counted how many times each motif occurred. Out of a total of >5,800 copies on chr14, 1,186 shared the most common motif **GGGG**TCAGAGGAATAGAAA**GGG**ACA**GGG**CTGAAGAACACAGGTCGCTGCATTTAGAAAGGAGGC, which was subsequently tested experimentally (see below).

Experimental validation of G4s in LSAU and Walusat. We aligned all G4 motifs overlapping with the LSAU motif to each other using mafft v7.481[98], and observed several patterns of G4s. On chromosomes 4 and 10,

there were many identical copies of several different motifs, while annotated LSAU regions on other chromosomes had more diverged sequences with many mismatches between the motifs. To group them together, we ran the sequence cluster algorithm starcode[99] separately on each strand. This clustered similar motifs and returned the consensus sequence. We then extracted the average methylation scores for all G4s in the top five clusters and visually inspected the distribution of these scores for each cluster. Three clusters that showed low methylation in combination with fairly high Quadron stability scores (Fig. S17) were selected for experimental validation. Two of them had uniform motifs (**GGGGG**C**GGGGGG**T**GGGGG**T**GGGG**A**GGGGG**CGGTCAGGCGGC**GGGG**T**GGG** with Quadron score 31.44, and **GGG**CGGCTGCA**GGGG**CCC**GGG**C**GGG**C**GGG**CGACGGTGGCGC**GGG** with Quadron score 19.76). The third cluster contained several very similar but not identical sequences, of which only one was chosen for validation (**GGG**T**GGGG**TGT**GGGGG**T**GGGG**A**GGGG**TGGTCAGGC**GGGGG**T**GGG**, Quadron score 31.01). Single-strand oligos were constructed from the above sequences and investigated by circular dichroism (CD), UV absorption spectra, and native polyacrylamide gel electrophoresis (PAGE), as described in[100]. In short, CD measurements were performed at 23°C and samples were measured in potassium ion only (110 mM K^+= 10mM K-phosphate, pH 7 + 95 mM KCl) in 1μM strand concentration, allowing 1 day to form due to many G-blocks. Isothermal difference spectra (IDS) were obtained by calculating the difference between the absorption spectra of the unfolded (1mM Na-phosphate) and folded (110mM K^+) forms of samples during an increase of ionic strength. Thermal difference spectra (TDS) were calculated as the difference of the unfolded (95°C) and folded (20°C) forms from temperature dependences in the potassium environment. Temperature dependencies were measured repeatedly (up-down-up-down) one day after K^+ addition. PAGE was run in 110 mM K^+ (10mM K-phosphate, pH 7 + 95 mM KCl) at 23°C. Samples were prepared either immediately (loaded on the gel after adding K^+) or 24 h before loading onto the gel.

Non-B distribution along the chromosomes and enrichment at centromeres. The density of each non-B motif type along the genome was calculated in 100-kb non-overlapping windows, generated with bedtools makewindows. For heatmap visualisation of acrocentric chromosomes and centromeres, such densities were normalized by the highest value to the scale from 0 to 1. Centromeric regions were taken from the GenomeFeatures tracks downloaded from https://www.genomeark.org/ for each species and converted to bed format. For chromosomes with two or more annotated active centromere ('CEN') regions with a satellite ('SAT') in between, we combined the active centromeres for the enrichment analysis without including the intermediate regions. Fold enrichment was calculated as non-B motif density within centromeres divided by the genome-wide non-B DNA density. No GC correction was performed as the centromeres are large regions with GC content very similar to the genome-wide average (Fig. S18). To test for significance of non-B DNA enrichment at the centromeres, the non-centromeric parts of each chromosome were divided into 100 windows with the same size as the actual centromere (for most chromosomes the windows had to overlap, however, if there were more than 100 possible non-overlapping windows, they were chosen randomly). Then, non-B DNA fold enrichment was calculated for each window separately, and the 100 values obtained for each chromosome were used as a null distribution to compare the centromere enrichment to. If the centromere fell outside the 0.025th and the 0.975th quantiles, the enrichment was considered to be significant. For detailed figures of centromeric and acrocentric regions, tracks with centromeric satellite repeats were downloaded from GenomeArk and added using the color scheme from UCSC genome browser. This included annotations of active α-sat (the parts that associate with the kinetochore proteins, usually the longest HOR array on each chromosome), inactive α-sat (HOR arrays that do not associate with the kinetochore), divergent αsat (older HORs that have started to erode), and monomeric α-sat (repeats not organized into HORs)[73]. CENP-B annotation files were downloaded from GenomeArk (for non-human apes) and from the supplementary database S15 in [73]. Suprachromosomal family information was extracted from the centromeric satellite annotation.

Circular density plots were generated with Circos[101]. Figures, as well as all statistical tests, were generated in R v4.4.0[102] using the tidyverse[103], ggupset[104], patchwork[105], cowplot[106], ggtext[107] and ggh4x[108] libraries.

Code availability

All code used for running our analyses and all in-house scripts generated for this paper are available on github: https://github.com/makovalab-psu/T2T_primate_nonB. Non-B DNA annotations are available at the UCSC

Genome Browser hub for the ape T2T genomes (https://github.com/marbl/T2T-Browser).

Acknowledgments

We are grateful to Glennis Logsdon, Karen Miga, Jennifer Gerton, Saswat Mohanty, Edmundo Torres-Gonzalez, Karol Pál, and Jacob Sieg for discussions of the results and useful suggestions. Saswat Mohanty provided the annotations of functional regions, based on a script written by Karol Pál. Robert Harris provided code for running winnowmap. Eddie Yong Hwee Loh provided useful explanations for methylation analysis. Glennis Logsdon and Karen Miga provided information on centromere annotations. Matthias Weissensteiner helped with generating Figure 1. This research was supported by the grant R35GM151945 and by the Willaman Chair Endowment Fund from the Eberly College of Science to KDM. Computations were performed at the Penn State Institute of Computational Data Sciences. This work was also supported by the grant 21–00580S from the Czech Science Foundation awarded to EK.

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. 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).
4. 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).
5. Zyner, K. G. *et al.* G-quadruplex DNA structures in human stem cells and differentiation. *Nat. Commun.* **13**, 142 (2022).
6. 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.
7. Cer, R. Z. *et al.* Non-B DB: a database of predicted non-B DNA-forming motifs in mammalian genomes. *Nucleic Acids Res* **39**, D383–91 (2011).
8. Prorok, P. *et al.* Involvement of G-quadruplex regions in mammalian replication origin activity. *Nat. Commun.* **10**, 3274 (2019).
9. Akerman, I. *et al.* A predictable conserved DNA base composition signature defines human core DNA replication origins. *Nat. Commun.* **11**, 4826 (2020).

10. 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).
11. Moye, A. L. *et al.* Telomeric G-quadruplexes are a substrate and site of localization for human telomerase. *Nat. Commun.* **6**, 7643 (2015).
12. 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).
13. Spiegel, J. *et al.* G-quadruplexes are transcription factor binding hubs in human chromatin. *Genome Biol.* **22**, 117 (2021).
14. 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).
15. Gazanion, E. *et al.* Genome wide distribution of G-quadruplexes and their impact on gene expression in malaria parasites. *PLoS Genet.* **16**, e1008917 (2020).
16. Saranathan, N. & Vivekanandan, P. G-Quadruplexes: More Than Just a Kink in Microbial Genomes. *Trends Microbiol.* **27**, 148–163 (2019).
17. 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).
18. 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).
19. 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).
20. 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).
21. 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).
22. 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).

23. 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).

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

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

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

27. 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).

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

29. 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).

30. 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).

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

32. 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).

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

34. 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.

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

36. Kipling, D. & Warburton, P. E. Centromeres, CENP-B and Tigger too. *Trends Genet* **13**, 141–145 (1997).

37. Goldberg, I. G., Sawhney, H., Pluta, A. F., Warburton, P. E. & Earnshaw, W. C. Surprising deficiency of CENP-B binding sites in African green monkey alpha-satellite DNA: implications for CENP-B function at centromeres. *Mol Cell Biol* **16**, 5156–5168 (1996).

38. Patchigolla, V. S. P. & Mellone, B. G. Enrichment of Non-B-Form DNA at D. melanogaster Centromeres. *Genome Biol Evol* **14**, (2022).

39. Liu, Q. *et al.* Non-B-form DNA tends to form in centromeric regions and has undergone changes in polyploid oat subgenomes. *Proc Natl Acad Sci U S A* **120**, e2211683120 (2023).

40. Yi, C. *et al.* Non-B-form DNA is associated with centromere stability in newly-formed polyploid wheat. *Sci China Life Sci* **67**, 1479–1488 (2024).

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

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

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

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

45. 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).

46. 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).

47. 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).

48. 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).

49. 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).

50. 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.

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

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

53. 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).

54. 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).

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

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

57. 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).

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

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

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

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

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

63. 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).

64. 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).
65. Sahakyan, A. B. *et al.* Machine learning model for sequence-driven DNA G-quadruplex formation. *Sci. Rep.* **7**, 14535 (2017).
66. SVA retrotransposons: Evolution and genetic instability. *Seminars in Cancer Biology* **20**, 234–245 (2010).
67. Hoyt, S. J. *et al.* From telomere to telomere: The transcriptional and epigenetic state of human repeat elements. *Science* **376**, eabk3112 (2022).
68. Gershman, A. *et al.* Epigenetic patterns in a complete human genome. *Science* **376**, eabj5089 (2022).
69. Halder, R. *et al.* Guanine quadruplex DNA structure restricts methylation of CpG dinucleotides genome-wide. *Mol. Biosyst.* **6**, 2439–2447 (2010).
70. Mao, S.-Q. *et al.* DNA G-quadruplex structures mold the DNA methylome. *Nat. Struct. Mol. Biol.* **25**, 951–957 (2018).
71. Gerton, J. L. A working model for the formation of Robertsonian chromosomes. *J Cell Sci* **137**, (2024).
72. de Lima, L. G. *et al.* The formation and propagation of human Robertsonian chromosomes. *bioRxiv* (2024) doi:10.1101/2024.09.24.614821.
73. Altemose, N. *et al.* Complete genomic and epigenetic maps of human centromeres. *Science* **376**, eabl4178 (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. Langley, S. A., Miga, K. H., Karpen, G. H. & Langley, C. H. Haplotypes spanning centromeric regions reveal persistence of large blocks of archaic DNA. *Elife* **8**, (2019).
76. Kejnovsky, E., Tokan, V. & Lexa, M. Transposable elements and G-quadruplexes. *Chromosome Res.* **23**, 615–623 (2015).
77. Kejnovsky, E. & Lexa, M. Quadruplex-forming DNA sequences spread by retrotransposons may serve as genome regulators. *Mob. Genet. Elements* **4**, e28084 (2014).
78. Lexa, M. *et al.* Guanine quadruplexes are formed by specific regions of human transposable elements.

BMC Genomics **15**, 1032 (2014).

79. Nicoletto, G. *et al.* G-quadruplexes in an SVA retrotransposon cause aberrant TAF1 gene expression in X-linked dystonia parkinsonism. *Nucleic Acids Res* **52**, 11571–11586 (2024).
80. Meneveri, R. *et al.* Molecular organization and chromosomal location of human GC-rich heterochromatic blocks. *Gene* **123**, 227–234 (1993).
81. Meneveri, R., Agresti, A., Rocchi, M., Marozzi, A. & Ginelli, E. Analysis of GC-rich repetitive nucleotide sequences in great apes. *J Mol Evol* **40**, 405–412 (1995).
82. Butterfield, R. J., Dunn, D. M., Duval, B., Moldt, S. & Weiss, R. B. Deciphering D4Z4 CpG methylation gradients in fascioscapulohumeral muscular dystrophy using nanopore sequencing. *Genome Res* **33**, 1439–1454 (2023).
83. Brázda, V. *et al.* G4Hunter web application: a web server for G-quadruplex prediction. *Bioinformatics* **35**, 3493–3495 (2019).
84. Bedrat, A., Lacroix, L. & Mergny, J.-L. Re-evaluation of G-quadruplex propensity with G4Hunter. *Nucleic Acids Res* **44**, 1746–1759 (2016).
85. Wang, G., Christensen, L. A. & Vasquez, K. M. Z-DNA-forming sequences generate large-scale deletions in mammalian cells. *Proc Natl Acad Sci U S A* **103**, 2677–2682 (2006).
86. Chittoor, S. S. & Giunta, S. Comparative analysis of predicted DNA secondary structures infers complex human centromere topology. *Am J Hum Genet* (2024) doi:10.1016/j.ajhg.2024.10.016.
87. Ohzeki, J.-I., Nakano, M., Okada, T. & Masumoto, H. CENP-B box is required for de novo centromere chromatin assembly on human alphoid DNA. *J Cell Biol* **159**, 765–775 (2002).
88. Sen Gupta, A. *et al.* Defining a core configuration for human centromeres during mitosis. *Nat Commun* **14**, 7947 (2023).
89. Brázda, V., Bartas, M. & Bowater, R. P. Evolution of Diverse Strategies for Promoter Regulation. *Trends Genet* **37**, 730–744 (2021).
90. Yella, V. R. & Vanaja, A. Computational analysis on the dissemination of non-B DNA structural motifs in promoter regions of 1180 cellular genomes. *Biochimie* **214**, 101–111 (2023).
91. Mohanty, S. K., Chiaromonte, F. & Makova, K. D. Evolutionary Dynamics of G-Quadruplexes in Human and Other Great Ape Telomere-to-Telomere Genomes. *bioRxiv* 2024.11.05.621973 (2024)

doi:10.1101/2024.11.05.621973.

92. Sinden, R. R., Pytlos-Sinden, M. J. & Potaman, V. N. Slipped strand DNA structures. *Front Biosci* **12**, 4788–4799 (2007).
93. Ma, H. *et al.* Centromere Plasticity With Evolutionary Conservation and Divergence Uncovered by Wheat 10+ Genomes. *Mol Biol Evol* **40**, (2023).
94. Jia, H. *et al.* Low-input PacBio sequencing generates high-quality individual fly genomes and characterizes mutational processes. *Nat Commun* **15**, 5644 (2024).
95. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
96. Jain, C. *et al.* Weighted minimizer sampling improves long read mapping. *Bioinformatics* **36**, i111–i118 (2020).
97. 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).
98. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**, 772–780 (2013).
99. Zorita, E., Cuscó, P. & Filion, G. J. Starcode: sequence clustering based on all-pairs search. *Bioinformatics* **31**, 1913–1919 (2015).
100. Kejnovská, I. *et al.* Clustered abasic lesions profoundly change the structure and stability of human telomeric G-quadruplexes. *Nucleic Acids Res* **45**, 4294–4305 (2017).
101. Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res* **19**, 1639–1645 (2009).
102. R Core Team. *R: A Language and Environment for Statistical Computing*. (Vienna, Austria , 2024).
103. Wickham, H. *et al.* Welcome to the Tidyverse. *Journal of Open Source Software* **4**, 1686 (2019).
104. Ahlmann-Eltze, C. *Ggupset: Combination Matrix Axis for 'ggplot2' to Create 'UpSet' Plots*. (2024).
105. Pedersen, T. L. *Patchwork: The Composer of Plots*. (2024).
106. Wilke, C. O. *Cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'*. (2024).
107. Claus O. Wilke, B. M. W. *Ggtext: Improved Text Rendering Support for 'ggplot2'*. (2024).
108. van den Brand, T. *ggh4x: Hacks for 'ggplot2'*. (2024).