Title: Non-B DNA affects polymerization speed and error rate in sequencers and living cells

Authors: Wilfried M. Guiblet[1#], Marzia A. Cremona[2#], Monika Cechova[3], Robert S. Harris[3], Iva Kejnovska[4], Eduard Kejnovsky[5], Kristin Eckert[6], Francesca Chiaromonte[2,7]*, and Kateryna D. Makova[3]*

Affiliations:
[1]Bioinformatics and Genomics Graduate Program, Penn State University, University Park, PA 16802 USA
[2]Department of Statistics, Penn State University, University Park, PA 16802 USA
[3]Department of Biology, Penn State University, University Park, PA 16802 USA
[4]Department of CD Spectroscopy of Nucleic Acids, Institute of Biophysics, Academy of Sciences of the Czech Republic, Brno, Czech Republic
[5]Department of Plant Developmental Genetics, Institute of Biophysics, Academy of Sciences of the Czech Republic, Brno, Czech Republic
[6]Department of Pathology, Penn State University, College of Medicine, Hershey, PA 17033 USA
[7]Sant'Anna School of Advanced Studies, Pisa, Italy

[#]These authors contributed equally
*Corresponding authors

ABSTRACT:

DNA conformation may deviate from the classical B-form in ~13% of the human genome. Non-B DNA regulates many cellular processes; however, its effects on DNA polymerization speed and accuracy have not been investigated genome-wide. Such an inquiry is critical for understanding neurological diseases and cancer genome instability. Here we present the first study of DNA polymerization kinetics in the human genome sequenced with Single-Molecule-Real-Time technology. We show that polymerization speed differs between non-B and B-DNA: it decelerates at G-quadruplexes and fluctuates periodically at disease-causing tandem repeats. We demonstrate that non-B DNA affects sequencing errors and human germline (1,000 Genomes Project, human-orangutan divergence, re-sequenced trios) and somatic (The Cancer Genome Atlas) mutations. Thus, non-B DNA has a large impact on genome evolution and human diseases.

MAIN TEXT:

The three-dimensional conformation of DNA at certain sequence motifs may deviate from the canonical double-stranded B-DNA (the right-handed helix with 10 nucleotides per turn) (*1*) in helix orientation and strand number (*2–4*). Approximately 13.2% of the human genome (394.2 Megabases) has the potential to form non-B DNA structures (Table S1), which are implicated in a myriad of cellular processes, and are associated with cancer and neurological diseases (*3–9*). For instance, adjacent runs of guanines can form G-quadruplex (G4) structures (Fig. 1A) (*10*) that participate in telomere maintenance (*11*), transcriptional regulation (*12*), and replication initiation (*13–15*). Consequently, G4 structures have emerged as attractive anti-cancer therapeutic targets (*16*). Additional non-B DNA structures (Fig. 1A) that are associated with transcriptional regulation include left-handed Z-DNA duplexes formed within alternating purine-pyrimidine sequences (*17*, *18*), A-phased repeats with helix bending formed within A-rich tracts (*19*, *20*), and H-DNA triplexes formed within polypurine/polypyrimidine tracts and mirror repeats (*21–23*). Finally, Short Tandem Repeats (STRs) (*24*), which also affect gene expression (*25*), can adopt slipped-strand (*26*) and other non-B DNA conformations (*27*). Expansions of STRs are associated with numerous neurological and muscular degenerative diseases (*8*, *28*, *29*). Notably, expansions of the hexanucleotide STR forming a G4 structure within the *C9orf72* gene is the most common genetic cause of amyotrophic lateral sclerosis (ALS) (*30*, *31*). Moreover, STRs are enriched in cancer-related genes and participate in their functions (*32*, *33*). Thus, growing evidence indicates that non-B DNA plays a pivotal role in several cellular pathways impacting health and disease.

Whereas the transient ability of non-B DNA motifs to form non-canonical structures regulates many cellular processes (*4*), these structures can also affect DNA synthesis and lead to genome instability, and thus can be viewed as both a blessing and a curse (*34*, *35*). *In vitro* and *ex vivo* studies of individual loci showed that non-B DNA formation inhibits prokaryotic

and eukaryotic DNA polymerases, causing their pausing and stalling of a replication fork (*36–42*). These processes have been postulated to underlie non-B DNA-induced genome instability, i.e. increase in chromosomal rearrangements, including those observed in cancer (*43–45*). Because the effect of non-B DNA on mutagenesis is driven by both the inherent DNA sequence and polymerase fidelity (*46*, *47*), we hypothesized that these structures can impact the efficiency and accuracy of DNA synthesis in sequencing instruments, as well as germline and somatic mutations in living cells — a possibility never examined previously. Despite the critical importance of non-B DNA structures, ours is the first genome-wide study of their impact on polymerization speed and mutation rates.

To evaluate whether DNA polymerization speed (i.e. polymerization kinetics) and polymerase errors are affected by non-B DNA, we utilized data from Single-Molecule Real Time (SMRT) sequencing. In addition to determining the primary nucleotide sequence, this technology, which uses an engineered bacteriophage phi29 polymerase (*48*), records Inter-Pulse Durations (IPDs; Fig. 1B), i.e the times between two fluorescent pulses corresponding to the incorporation of two consecutive nucleotides (*49*). We use IPDs as a measure of polymerization kinetics. Compared with other methods (e.g., (*50*)), SMRT sequencing is presently the only high-throughput technology allowing a direct, simultaneous investigation of the genome-wide effects of several non-B DNA motif types on polymerization kinetics.

Polymerization kinetics at non-B DNA motifs

To conduct such an investigation, we considered 92 different motif types potentially forming non-B DNA (Fig. 1A; Tables S2-S3) (*4*), retrieving positions of predicted motifs from the non-B DNA DataBase (*51*) and annotating STRs (*52*). We constructed motif-containing genomic windows taking ±50 bp from the center of each motif (most were shorter than 100 bp; Fig. S1) and excluded windows comprising multiple motifs (Tables S2-S3). For controls, we constructed 100-bp motif-free windows to represent genomic background, i.e. putative B-

DNA. We populated each motif-containing and motif-free window with 100 single-nucleotide resolution IPDs (Fig. 1B) from the genome of a human male previously sequenced with SMRT at 69x (*53*). This was performed separately for the reference and reverse complement strands, because each strand is used separately as a template during SMRT sequencing (Fig. 1B). For each motif type, we then aligned the centers of all motifs and aggregated IPD curves across windows, producing a distribution of IPD curves per motif type, per strand (Fig. 1B).

To evaluate whether non-B motifs present polymerization kinetics patterns different from B-DNA, we used Interval-Wise Testing (*54*), a novel Functional Data Analysis (FDA) approach (*55*) to identify genomic sites (i.e. bases) or intervals where IPD curve distributions significantly differ between motif-containing and motif-free 100-bp windows (Fig. 2A-E; two-sided test, see Methods). We indeed found altered polymerization kinetics in and/or around several non-B DNA motifs. Below, we describe results for the reference strand (a total of 2,916,328 motif-containing and 2,524,489 motif-free windows; upper panels in Fig. 2A-D; Fig. 2E; Figs. S2-S10, S11A, S11C, S11E); results for the reverse complement serve as a biological replicate (lower panels in Fig. 2A-D; Fig. S11B, S11D, S11F).

Two lines of evidence are consistent with G4 motifs hindering polymerase progression. First, they decreased polymerization speed. Compared to motif-free windows, G4-containing windows had significantly higher IPDs near their centers, i.e. near the motifs (up to 1.7-fold IPD increase at the 95th quantile; Fig. 2A). All G4 motif types exhibited this elevation, although the IPD curve shapes differed depending on the G4 motif sequence (Fig. S3). Furthermore, the shape of the IPD distribution encompassing all G4 motif types remained the same (Fig. S4) when we limited our analysis to G4 motifs forming the most stable G4 quadruplexes, as identified by *in vitro* ion concentration manipulations (*50*). Second, sequencing depth was lower at G4 motifs than at motif-free windows (86% of motif-free depth; Fig. 2A), suggesting that the former interfere with polymerization. Polymerization slowdown and decreased sequencing depth were evident on the reference strand where

G4s were annotated (upper panel in Fig. 2A; “G4+” in Fig. 2E), consistent with G-quadruplex structures forming only on the guanine-rich strand (*6*). Notably, elevated IPDs were observed in all sequencing passes through the same G4+ containing circular template (Figs. 1B and S5), suggesting that the structure is not resolved during sequencing. In contrast, the corresponding opposite strand (lower panel in Fig. 2A), as well as the reference strand where G4s were annotated on the reverse complement strand (“G4-” in Fig. 2E), — both cytosine-rich — showed a significant overall polymerization acceleration and displayed a smaller decrease in sequencing depth (92% of motif-free depth).

We observed that several other non-B DNA motifs significantly altered polymerization kinetics — e.g., A-phased repeats, inverted repeats, mirror repeats, and Z-DNA (Figs. 2E and S6). In contrast to the G4 motifs, the effects on polymerization kinetics were similar on the two sequenced strands (Figs. 2E and S11B), suggesting that either both strands are required for non-B DNA formation at these motifs or non-B DNA can be formed with similar probability on each strand (Fig. 1B).

Additionally, we found that STRs altered polymerization kinetics in length- and sequence-dependent manner (Figs. S8E-F, 2B-E, S7-S10); these variables impact the types and stability of non-B DNA structures that can form in addition to slipped structures (*24*). For STRs with ≥2-nucleotide repeated units, the variation in polymerization kinetics was periodic, with the period (in bases) matching the length of the repeated unit, consistent with effects of strand slippage (Figs. 2B-E, S7-S10). This pattern was evident for trinucleotide STRs whose length variants at some loci are associated with neurological diseases (Fig. 2B-D), e.g., $(CGG)_n$ implicated in Fragile X syndrome, $(CAG)_n$ implicated in Huntington’s disease and Spinocerebellar ataxia, and $(GAA)_n$ implicated in Friedreich’s ataxia (*28*). Genome-wide, $(CGG)_n$ repeats showed a strong periodic decrease in polymerization speed (elevated IPDs) on the annotated strand (up to 9-fold IPD increase at the 95th quantile; Fig. 2C), consistent with their ability to form G4-like structures and hairpins (*56*), and corroborating an analysis of

27 $(CGG)_n$ occurrences in the *E. coli* genome (*57*). The pattern for $(CAG)_n$ repeats, also capable of forming hairpins (*27*), was similar (Fig. 2B). Globally, STRs capable of forming hairpins (Table S4) presented the most striking polymerization deceleration and periodicity (Figs. 2B-C, 2E and S7). In contrast, STRs forming H-DNA (Table S4), including $(GAA)_n$, accelerated polymerization (lowered IPDs; Figs. 2D-E and S8). For many STRs, significant deviations from background IPD levels were shifted 5' to the annotated motif (Figs. 2E and S11), possibly due to polymerase stalling caused by difficulty in accommodating the alternative DNA structure within the polymerase active site.

We examined whether the alterations in polymerization kinetics that we observed at non-B DNA motifs could be explained by (1) base modifications, or (2) single or di- nucleotide composition. First, we observed that IPD patterns for most non-B DNA motifs were still clearly detectable in amplified DNA (Fig. S12), suggesting that they were not due to base modifications in the original template DNA (*49*). Second, we found that the mean IPD in motif-free windows (IPDs averaged across 100 nucleotides for each window) depended significantly on nucleotide composition ($p<2\times10^{-16}$; compositional regression models (*58*) with single nucleotides or with dinucleotides). Considering single nucleotide composition, polymerization decelerated with an increased proportions of guanines, and accelerated with an increased proportion of thymines, but was not significantly affected by the proportions of adenines or cytosines (Fig. S13). However, compositional regressions with either single nucleotide or dinucleotide composition explained only a relatively small portion of mean IPD variation among motif-free windows; 11.5% for single nucleotides and 20.8% for dinucleotides (similar to results obtained for bacterial genomes (*59*)). Moreover, the mean IPDs in most motif-containing windows were significantly higher or lower (e.g., G4+ and G4- containing windows, respectively) than those predicted by such regression (Figs. 2F and S14). Altogether, nucleotide composition falls far short of explaining IPD variations at non-B DNA motifs. In particular, the presence of guanines in G4+ motifs cannot explain the overall substantial deceleration of polymerization observed at these sites.

Polymerization kinetics and biophysical characteristics of G-quadruplexes

To experimentally test whether non-B DNA structures can form at predicted motifs, we investigated the relationship between polymerization kinetics and biophysical characteristics of the ten G4 motifs most common in the human genome (Table S5). According to circular dichroism spectroscopy (CD) and native polyacrylamide gel electrophoresis (PAGE) analyses, all ten motifs quickly formed stable quadruplexes at low potassium concentrations, suggesting that they have a high propensity to form such structures (*60*) albeit with different molecularity (intra- or intermolecular) and strand orientations (parallel or antiparallel, Table S5). Using regressions for intramolecular G4s, we found a significant positive relationship between mean IPD and delta epsilon (Fig. 3A; $p<2\times10^{-16}$, R-squared=32.3%), a measure of structure organization quality obtained by CD, and between mean IPD and melting temperature (Fig. 3B; $p<2\times10^{-16}$, R-squared=5.7%), a measure of thermostability and structure denaturation obtained by light absorption (Table S5; results for intermolecular G4s are shown in Fig. S15) (*60*). Thus, polymerization slowdown and biophysical characteristics of G4 formation correlate, strongly suggesting that the motifs indeed form G4 structures during the SMRT sequencing reaction (intramolecular G4 structures are only a few nanometers in diameter (*61*) and thus can fit within the 60x100 nm wells of Pacific Biosciences, or PacBio, instruments (*62*)).

While not possessing a canonical G4 motif, the $(GGT)_n$ STR has an IPD profile similar to that of G4+ (Figs. 2E and S10E) and its reverse complement $(ACC)_n$ has a profile similar to that of G4- (Figs. 2E and S10B), suggesting that $(GGT)_n$ may fold into a G4-like structure. Remarkably, biophysical analyses (CD, native PAGE, and thermal denaturation) also showed that $(GGT)_n$ motifs indeed adopt quadruplex conformation (Fig. S16; Table S6). Thus, we have evidence that statistical FDA techniques applied to polymerization kinetics data can enable non-B DNA structure discovery.

Sequencing error rates at non-B DNA

To examine whether phi29 polymerase accuracy is affected during synthesis of non-B DNA motifs in the genome, we contrasted SMRT sequencing error rates between such motifs and motif-free regions, using the same data employed to study polymerization kinetics — i.e. the genome of a human male sequenced with SMRT at 69x (*53*). We focused on motifs themselves (as opposed to 100-bp motif-containing windows), and for controls, identified motif-free regions matched to motifs in number and length. Because of the potential for inaccurate typing of STRs (*52*) and for motif misalignments in repetitive loci, we restricted our attention to six non-STR motif types present on the reference strand of the non-repetitive portion of the genome (Fig. 4; Table S7-S8). We also excluded fixed differences between sequenced (*53*) and reference genomes, and computed error rates as the proportion of variants (relative to hg19) within the total number of nucleotides sequenced for the motif or motif-free region — including errors supported even by a single read (see Methods). Below we present results for errors on the newly synthesized strand that uses the template strand annotated with non-B DNA motifs. We observed a strong effect of G4 motifs on SMRT error rates. Mismatches were strongly elevated on the newly synthesized strand when G4s were present on the template strand (i.e. when G4s were annotated on the reference strand; Fig. 4A; 1.75-fold elevation in G4+). Deletions were increased in both G4+ and G4- (Fig. 4B; 1.46- and 1.08-fold, respectively). Insertions, the most common error type for SMRT sequencing, were depressed when G4+ and particularly G- were used as templates (Fig. S17; 1.03- and 1.23-fold, respectively). Among other motif types with sizeable effects, Z-DNA displayed depressed mismatches (1.19-fold; Fig. 4A) and deletions (1.17-fold; Fig. 4B), but increased insertions (1.16-fold; Fig. S17). In summary, SMRT sequencing errors of all three types had different rates in non-B motifs compared to motif-free regions, with strong elevations of mismatches and deletions at G4- motifs (Fig. 4).

To assess whether non-B DNA affects error rates of polymerases others than the phi29 polymerase used in SMRT sequencing (*48*), we repeated our analysis for high-depth Illumina sequencing data, generated with the Pyrococcus-derived Phusion polymerase (*63*), for the same human whose genome was sequenced with SMRT (*53*). We again restricted attention to the non-repetitive portion of the genome, and removed motifs and motif-free regions harboring fixed differences. We also performed a simulation study (Supplementary Note 1) and a comparison of aligners (Supplementary Note 2) to verify that our results were not driven by misalignments or the use of specific alignment software. We present results for errors on the newly synthesized strand that uses the template strand annotated with non-B DNA motifs, and only for Read 1 (Figs. 4 and S17; the error analysis for Read 2 and for the other strand, as well as for overall Illumina errors at non-B motifs, is discussed in Supplementary Note 3). Similar to SMRT, Illumina displayed increased mismatches in G4+ (2.71-fold, respectively) and increased deletions in G4+ and G4- (3.45- and 4.77-fold, respectively; Fig. 4). In contrast to SMRT, Illumina mismatch errors were also elevated in G4- (2.73-fold) and Z-DNA (1.53-fold), and reduced at A-phased repeats (1.20-fold). Moreover, mismatches and deletions were *strongly* elevated in direct repeats (4.29- and 2.78-fold, respectively; Fig. 4). Notwithstanding these polymerase-/technology-specific differences, our results demonstrate that error rates of two different polymerases — phi29 (SMRT) and Phusion (Illumina) — are affected by the presence of several non-B DNA motif types.

Mutation rates at non-B DNA

Mutation rates are known to be non-uniform across the genome (*64*); however, the mechanisms leading to such regional variation are not yet entirely understood (*65*). Our results on sequencing errors for two different polymerases/technologies, as well as previous *in vitro* polymerase studies (Usdin 1995; Kang 1995; Hile and Eckert 2004; Delagoutte 2008; Eddy 2015; Barnes 2017) demonstrating the effects of non-B DNA on progression of phage,

prokaryotic and eukaryotic polymerases, raise an intriguing question: are germline and/or somatic mutation rates *in vivo* also affected by these motifs? To date, this question has not been addressed on a genome-wide scale. To do so, we used four data sets that capture mutations arising in living cells: human-orangutan divergence (*66*), 1,000 Genomes Project variants (*67*), *de novo* mutations in Icelandic trios (*74*) (proxies for germline mutations), and mutations from The Cancer Genome Atlas (*68*) (TCGA, a proxy for somatic mutations). Note that in these data sets we cannot differentiate the strand where a mutation occurred, thus, for instance, results are reported for G4+ and G4- containing strands combined.

We first tested whether the number of nucleotide substitutions, insertions and deletions in human-orangutan genomic alignments (*66*) differ between non-B DNA motifs and motif-free regions matched to motifs in number, length, and broad genomic location (to account for megabase-scale variation in mutation rates across the genome, see Methods) (*64*, *69*). Since human and orangutan reference genomes were generated with the more accurate Sanger sequencing (which uses a T7-derived polymerase (*70*)), the vast majority of differences between them should result from germline mutations — not sequencing errors. Nucleotide substitutions were significantly elevated in G4s (Fig. 4A; 1.15-fold) and Z-DNA (1.78-fold), and depressed in A-phased and direct repeats (1.11- and 1.59-fold, respectively). Elevated point mutations in Z-DNA were previously observed in plasmid reporter assays (*71*). Fixed deletions were elevated in mirror repeats (1.20-fold; Fig. 4B). Fixed insertions were significantly elevated in G4s (Fig. S17; 3.75-fold) and inverted (1.54-fold), mirror (1.43-fold), and particularly direct (11.69-fold) repeats. Elevation of insertions in direct repeats is consistent with the DNA slippage mechanism of insertions (*72*).

These divergence-based results may underestimate the true effects of non-B DNA motifs on germline mutations, due to potential purifying selection acting to conserve sequence and preserve function of some such motifs (*4*). To further examine whether non-B DNA motifs influence germline mutations, we analyzed mutations less affected by selection — Single

Nucleotide Polymorphisms (SNPs) and polymorphic insertions and deletions annotated by the 1000 Genomes Project (*67*). For this analysis, we focused on high-frequency variants (global minor allele frequency across all individuals ≥5%), because these are minimally affected by DNA damage (*73*) and extremely unlikely to be caused by Illumina sequencing errors (Supplementary Note 3). Results for the 1000 Genomes data did not change qualitatively if we analyzed variants with 1-5% allele frequency, which are less affected by selection but more affected by DNA damage (*73*) (Fig. S18), and largely mirrored those obtained for divergence data from human-orangutan alignments (Figs. 4 and S17). Specifically, we observed elevated SNPs in G4s (Fig. 4A; 1.30-fold) and Z-DNA (1.94-fold), and elevated deletions in inverted repeats (1.39-fold; Fig. 1B). Polymorphic insertions displayed trends similar to those observed for fixed insertions (Extended Data File 1 and Fig. S17, respectively), but the former were based on a smaller number of observations and were not significant.

Ideally, to examine mutation rates at non-B DNA, one would like to utilize *de novo* mutations inferred from resequencing of trios, the data on which are still rather limited to date. Examining the largest data set available, which consists of 1,548 Icelandic trios (*74*), we found elevated mismatches at G4s and decreased mismatches at A-phased and direct repeats (Supplementary Note 5) — consistent with our observations for diversity and divergence (Fig. 4). However, these results were not statistically significant because of the small number of mutations in the trio data (Extended Data File 1). In summary, although with varying degree of statistical confidence, four separate lines of evidence leveraging different evolutionary distances and sequencing technologies (Sanger sequencing for human and orangutan reference genomes (*66*, *75*); Illumina for the 1,000 Genomes Project (*67*) and for trio resequencing (*74*)) lend support to the notion that non-B motifs alter germline mutation rates.

Finally, we evaluated the effects of non-B DNA motifs on the rates of cancer somatic mutations using data from The Cancer Genome Atlas (*68*). The number of somatic SNPs per site (see Methods) was elevated in G4s Z-DNA (Fig. 4A; 2.55-fold), and depressed in A-phased, direct and inverted repeats (1.42-, 2.27-, 1.18-fold, respectively). Somatic deletions were elevated in G4s (2.53-fold), and inverted and mirror repeats (1.39- and 1.31-fold; Fig. 4B). The similarities observed between the effects on non-B DNA on somatic and germline mutations are suggestive of common mechanisms.

Discussion

Our genome-wide study demonstrates that SMRT sequencing polymerization kinetics is significantly altered at genome sequences capable of forming non-B DNA, with striking patterns for G4s and STRs. Importantly, we also demonstrate that analyzing polymerization kinetics data with FDA statistical techniques can enable non-B DNA structure discovery. We identified $(GGT)_n$ motifs as potentially forming a G4-like structure based on their polymerization pattern, and used biophysical profiling to validate the formation of such structure. With the increasing popularity of SMRT sequencing and growing publicly available data, we expect that our understanding and use of polymerization kinetics will expand rapidly. By analyzing hundreds of thousands of non-B DNA motifs, we observed polymerization slowdown and acceleration during SMRT sequencing for the majority of motifs analyzed. These results suggest the intriguing possibility that non-B DNA may act as a polymerization speed modifier in the genome also under natural conditions (*57*). The elevated sequence polymorphism of some non-B DNA motifs (e.g., G4-quadruplexes and Z-DNA) in populations, which is demonstrated by our results, could in fact induce interindividual differences in polymerization kinetics and genome instability. Also, our results lend support to the possibility that altered, periodic polymerization kinetics patterns at disease-associated STRs contribute to their instability (*76*).

We documented strong effects of non-B DNA motifs on sequencing errors. Elevated SMRT and Illumina errors in many non-B DNA motifs should be taken into account when evaluating sequencing results. Many such errors are likely corrected via deep or circular sequencing, but biases in the sequence consensus might remain — a possibility that needs to be evaluated in future studies. Interestingly, the decreased fidelity of the Phusion polymerase (used by Illumina technology) at G4 structures was noted in another recent study (*50*). In our study, the error frequencies of the two sequencing technologies examined were both influenced by the presence of non-B DNA.

Our analyses of primate divergence, human diversity, resequenced trios, and cancer somatic variation datasets further suggest a universally significant effect of non-B DNA on mutation rates. This was surprising, given the different biochemistry and fidelity of the DNA polymerases involved in sequencing (e.g., engineered phi29 polymerase for SMRT (*48*); Phusion for Illumina (*63*); T7-derived Sequenase for Sanger sequencing (*70*)) and in cell replication (*77*). Perhaps the differences in the magnitude of the non-B DNA effects that we observed for the various datasets (Fig. 4) reflect such differences in polymerase identity and fidelity. Moreover, some effects emerged only from diversity and divergence data, suggesting that mutagenesis at particular non-B structures *in vivo* arises by specific mechanisms. For instance, whereas SNPs and nucleotide substitutions were elevated in Z-DNA (and more strongly so than Illumina errors), potentially reflecting its sensitivity to genotoxic agents (*7*), PacBio sequencing nucleotide mismatch errors were depressed in these motifs.

Overall, our results suggest that non-B DNA is an important factor contributing to localized variation in mutation rates across the genome. Its effects occur at the scale of tens of nucleotides, and can be fairly large — e.g., based on divergence data, we observe 1.78- and 3.75-fold increases (relative to genomic background) for the rates of nucleotide substitutions in Z-DNA and of insertions in G4s, respectively. However, we also observe smaller effects — e.g., for nucleotide substitutions at G-quadruplexes and A-phased repeats (1.15-fold

increase and 1.11-fold decrease, respectively). Interestingly, these smaller effects match the magnitude of megabase-scale variation in substitution and indel rates across the genome: substitution rates computed from human-mouse alignments at a 5-Mb scale vary up to ~1.15-fold (*78*), and those computed from human-orangutan alignments at a 1-Mb scale up to ~1.10-1.33-fold (*64*), relative to the corresponding genome-wide averages.

We found elevated diversity and divergence in many non-B DNA motifs, despite their potential functional role and evolutionary constraint (*79*, *80*). Our findings, together with observations on the transient nature of non-B DNA conformations (*4*), portray non-B DNA as an effective, environmentally sensitive and fast modulator of genome structure — affecting a number of cellular processes. This is particularly intriguing in view of recent evidence broadening the spectrum of mechanisms through which non-B DNA may modulate the cell, encompassing, e.g., epigenetic instability (*34*), transposon silencing (*81*), and non-coding RNA regulation (*82*).

REFERENCES

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

2. S. M. Mirkin, Discovery of alternative DNA structures: a heroic decade (1979-1989). *Front. Biosci.* **13**, 1064–1071 (2008).

3. A. Bacolla, R. D. Wells, Non-B DNA conformations, genomic rearrangements, and human disease. *J. Biol. Chem.* **279**, 47411–47414 (2004).

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

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

6. N. Maizels, G4-associated human diseases. *EMBO Rep.*, e201540607 (2015).

7. G. Wang, K. M. Vasquez, Z-DNA, an active element in the genome. *Front. Biosci.* **12**, 4424–4438 (2007).

8. H. T. Orr, H. Y. Zoghbi, Trinucleotide repeat disorders. *Annu. Rev. Neurosci.* **30**, 575–621 (2007).

9. S. M. Mirkin, Expandable DNA repeats and human disease. *Nature*. **447**, 932–940 (2007).

10. D. Sen, W. Gilbert, Formation of parallel four-stranded complexes by guanine-rich motifs in DNA and its implications for meiosis. *Nature*. **334**, 364–366 (1988).

11. G. N. Parkinson, M. P. H. Lee, S. Neidle, Crystal structure of parallel quadruplexes from human telomeric DNA. *Nature*. **417**, 876–880 (2002).

12. A. Siddiqui-Jain, C. L. Grand, D. J. Bearss, L. H. Hurley, Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress c-MYC transcription. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 11593–11598 (2002).

13. A. K. Todd, M. Johnston, S. Neidle, Highly prevalent putative quadruplex sequence motifs in human DNA. *Nucleic Acids Res.* **33**, 2901–2907 (2005).

14. J. L. Huppert, S. Balasubramanian, Prevalence of quadruplexes in the human genome. *Nucleic Acids Res.* **33**, 2908–2916 (2005).

15. E. Besnard *et al.*, Unraveling cell type–specific and reprogrammable human replication origin signatures associated with G-quadruplex consensus motifs. *Nat. Struct. Mol. Biol.* **19**, nsmb.2339 (2012).

16. S. Balasubramanian, L. H. Hurley, S. Neidle, Targeting G-quadruplexes in gene promoters: a novel anticancer strategy? *Nat. Rev. Drug Discov.* **10**, 261–275 (2011).

17. A. H. Wang *et al.*, Molecular structure of a left-handed double helical DNA fragment at atomic resolution. *Nature*. **282**, 680–686 (1979).

18. B. Wittig, T. Dorbic, A. Rich, Transcription is associated with Z-DNA formation in metabolically active permeabilized mammalian cell nuclei. *Proc. Natl. Acad. Sci. U. S. A.*

88, 2259–2263 (1991).

19. H.-S. Koo, H.-M. Wu, D. M. Crothers, DNA bending at adenine\textperiodcentered thymine tracts. *Nature*. **320**, 501–506 (1986).

20. A. Jansen, E. van der Zande, W. Meert, G. R. Fink, K. J. Verstrepen, Distal chromatin structure influences local nucleosome positions and gene expression. *Nucleic Acids Res.* **40**, 3870–3885 (2012).

21. B. P. Belotserkovskii *et al.*, Mechanisms and implications of transcription blockage by guanine-rich DNA sequences. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 12816–12821 (2010).

22. G. P. Schroth, P. S. Ho, Occurrence of potential cruciform and H-DNA forming sequences in genomic DNA. *Nucleic Acids Res.* **23**, 1977–1983 (1995).

23. S. M. Mirkin *et al.*, DNA H form requires a homopurine–homopyrimidine mirror repeat. *Nature*. **330**, 495–497 (1987).

24. H. Ellegren, Microsatellites: simple sequences with complex evolution. *Nat. Rev. Genet.* **5**, 435–445 (2004).

25. S. Sawaya *et al.*, Microsatellite tandem repeats are abundant in human promoters and are associated with regulatory elements. *PLoS One*. **8**, e54710 (2013).

26. R. R. Sinden, Richard, R. Sinden, Slipped strand DNA structures. *Front. Biosci.* **12**, 4788 (2007).

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

28. A. L. Castel, J. D. Cleary, C. E. Pearson, Repeat instability as the basis for human diseases and as a potential target for therapy. *Nat. Rev. Mol. Cell Biol.* **11**, 165–170 (2010).

29. C. T. Caskey, A. Pizzuti, Y. H. Fu, R. G. Fenwick Jr, D. L. Nelson, Triplet repeat mutations in human disease. *Science*. **256**, 784–789 (1992).

30. A. E. Renton *et al.*, A hexanucleotide repeat expansion in C9ORF72 is the cause of chromosome 9p21-linked ALS-FTD. *Neuron*. **72**, 257–268 (2011).

31. M. DeJesus-Hernandez *et al.*, Expanded GGGGCC hexanucleotide repeat in noncoding region of C9ORF72 causes chromosome 9p-linked FTD and ALS. *Neuron*. **72**, 245–256 (2011).

32. Y. Haberman, N. Amariglio, G. Rechavi, E. Eisenberg, Trinucleotide repeats are prevalent among cancer-related genes. *Trends Genet.* **24**, 14–18 (2008).

33. B. E. Madsen, P. Villesen, C. Wiuf, Short tandem repeats and genetic variation. *Methods Mol. Biol.* **628**, 297–306 (2010).

34. A.-L. Valton, M.-N. Prioleau, G-Quadruplexes in DNA Replication: A Problem or a Necessity? *Trends Genet.* **32**, 697–706 (2016).

35. M. Tarsounas, M. Tijsterman, Genomes and G-quadruplexes: for better or for worse. *J. Mol. Biol.* **425**, 4782–4789 (2013).

36. G. M. Samadashwily, G. Raca, S. M. Mirkin, Trinucleotide repeats affect DNA

replication in vivo. *Nat. Genet.* **17**, 298–304 (1997).

37. M. M. Krasilnikova, S. M. Mirkin, Replication stalling at Friedreich's ataxia (GAA)n repeats in vivo. *Mol. Cell. Biol.* **24**, 2286–2295 (2004).

38. S. Eddy *et al.*, Evidence for the kinetic partitioning of polymerase activity on G-quadruplex DNA. *Biochemistry*. **54**, 3218–3230 (2015).

39. S. Kang, K. Ohshima, M. Shimizu, S. Amirhaeri, R. D. Wells, Pausing of DNA synthesis in vitro at specific loci in CTG and CGG triplet repeats from human hereditary disease genes. *J. Biol. Chem.* **270**, 27014–27021 (1995).

40. S. E. Hile, K. A. Eckert, Positive correlation between DNA polymerase alpha-primase pausing and mutagenesis within polypyrimidine/polypurine microsatellite sequences. *J. Mol. Biol.* **335**, 745–759 (2004).

41. R. P. Barnes, S. E. Hile, M. Y. Lee, K. A. Eckert, DNA polymerases eta and kappa exchange with the polymerase delta holoenzyme to complete common fragile site synthesis. *DNA Repair* . **57**, 1–11 (2017).

42. I. Voineagu, V. Narayanan, K. S. Lobachev, S. M. Mirkin, Replication stalling at unstable inverted repeats: interplay between DNA hairpins and fork stabilizing proteins. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 9936–9941 (2008).

43. G. Wang, K. M. Vasquez, Naturally occurring H-DNA-forming sequences are mutagenic in mammalian cells. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 13448–13453 (2004).

44. A. Bacolla *et al.*, Breakpoints of gross deletions coincide with non-B DNA conformations. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 14162–14167 (2004).

45. G. Wang, S. Carbajal, J. Vijg, J. DiGiovanni, K. M. Vasquez, DNA structure-induced genomic instability in vivo. *J. Natl. Cancer Inst.* **100**, 1815–1817 (2008).

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

47. G. Ananda *et al.*, Microsatellite interruptions stabilize primate genomes and exist as population-specific single nucleotide polymorphisms within individual human genomes. *PLoS Genet.* **10**, e1004498 (2014).

48. J. Eid *et al.*, Real-time DNA sequencing from single polymerase molecules. *Science*. **323**, 133–138 (2009).

49. B. A. Flusberg *et al.*, Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat. Methods*. **7**, 461–465 (2010).

50. V. S. Chambers *et al.*, High-throughput sequencing of DNA G-quadruplex structures in the human genome. *Nat. Biotechnol.* **33**, 877–881 (2015).

51. R. Z. Cer *et al.*, Non-B DB v2.0: a database of predicted non-B DNA-forming motifs and its associated tools. *Nucleic Acids Res.* **41**, D94–D100 (2012).

52. A. Fungtammasan *et al.*, Accurate typing of short tandem repeats from genome-wide sequencing data and its applications. *Genome Res.* **25**, 736–749 (2015).

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

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

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

56. Y. Nadel, P. Weisman-Shomer, M. Fry, The fragile X syndrome single strand d(CGG)n nucleotide repeats readily fold back to form unimolecular hairpin structures. *J. Biol. Chem.* **270**, 28970–28977 (1995).

57. S. Sawaya, J. Boocock, M. A. Black, N. J. Gemmell, Exploring possible DNA structures in real-time polymerase kinetics using Pacific Biosciences sequencer data. *BMC Bioinformatics*. **16**, 21 (2015).

58. J. Aitchison, *The statistical analysis of compositional data* (Chapman and Hall, 1986).

59. E. E. Schadt *et al.*, Modeling kinetic rate variation in third generation DNA sequencing data to detect putative modifications to DNA bases. *Genome Res.* **23**, 129–141 (2013).

60. J. Kypr, I. Kejnovská, D. Renciuk, M. Vorlícková, Circular dichroism and conformational polymorphism of DNA. *Nucleic Acids Res.* **37**, 1713–1725 (2009).

61. S. Neidle, S. Balasubramanian, Quadruplex Nucleic Acids, Royal Society of Chemistry (2006).

62. S. Turner *et al.*, Nanoscale apertures having islands of functionality. *US Patent* (2017), (available at https://www.google.com/patents/US9637380).

63. M. A. Quail *et al.*, Optimal enzymes for amplifying sequencing libraries. *Nat. Methods*. **9**, 10–11 (2011).

64. P. Kuruppumullage Don, G. Ananda, F. Chiaromonte, K. D. Makova, Segmenting the human genome based on states of neutral genetic divergence. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 14699–14704 (2013).

65. K. D. Makova, R. C. Hardison, The effects of chromatin organization on variation in mutation rates in the genome. *Nat. Rev. Genet.* **16**, 213–223 (2015).

66. D. P. Locke *et al.*, Comparative and demographic analysis of orang-utan genomes. *Nature*. **469**, 529–533 (2011).

67. 1000 Genomes Project Consortium *et al.*, A global reference for human genetic variation. *Nature*. **526**, 68–74 (2015).

68. The Cancer Genome Atlas Home Page. *The Cancer Genome Atlas - National Cancer Institute*, (available at http://cancergenome.nih.gov/).

69. A. Hodgkinson, A. Eyre-Walker, Variation in the mutation rate across mammalian genomes. *Nat. Rev. Genet.* **12**, 756–766 (2011).

70. B. Zhu, Bacteriophage T7 DNA polymerase - sequenase. *Front. Microbiol.* **5**, 181 (2014).

71. A. Bacolla *et al.*, Non-B DNA-forming sequences and WRN deficiency independently increase the frequency of base substitution in human cells. *J. Biol. Chem.* **286**, 10017–10026 (2011).

72. P. W. Messer, P. F. Arndt, The majority of recent short DNA insertions in the human

genome are tandem duplications. *Mol. Biol. Evol.* **24**, 1190–1197 (2007).

73. L. Chen, P. Liu, T. C. Evans Jr, L. M. Ettwiller, DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification. *Science*. **355**, 752–756 (2017).

74. H. Jónsson *et al.*, Whole genome characterization of sequence diversity of 15,220 Icelanders. *Sci Data*. **4**, 170115 (2017).

75. E. S. Lander *et al.*, Initial sequencing and analysis of the human genome. *Nature*. **409**, 860–921 (2001).

76. E. W. Loomis *et al.*, Sequencing the unsequenceable: expanded CGG-repeat alleles of the fragile X gene. *Genome Res.* **23**, 121–128 (2013).

77. L. A. Loeb, R. J. Monnat Jr, DNA polymerases and human disease. *Nat. Rev. Genet.* **9**, 594–604 (2008).

78. R. C. Hardison *et al.*, Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. *Genome Res.* **13**, 13–26 (2003).

79. R. Rohs *et al.*, The role of DNA shape in protein-DNA recognition. *Nature*. **461**, 1248–1253 (2009).

80. S. C. J. Parker, L. Hansen, H. O. Abaan, T. D. Tullius, E. H. Margulies, Local DNA topography correlates with functional noncoding regions of the human genome. *Science*. **324**, 389–392 (2009).

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

82. R. Simone, P. Fratta, S. Neidle, G. N. Parkinson, A. M. Isaacs, G-quadruplexes: Emerging roles in neurodegenerative diseases and the non-coding transcriptome. *FEBS Lett.* **589**, 1653–1668 (2015).

ACKNOWLEDGEMENTS:

We are grateful to Jonas Korlach and Sarah Kingan from Pacific Biosciences Inc. for comments on the manuscript, Rahul Vegesna for help with PacBio data formats, Fabio Cumbo for help with TCGA data formats, Marta Tomaszkiewicz for conducting whole-genome amplification, and Malcolm Ferguson-Smith for flow-sorting. Funding for the project was provided by the Huck Institutes for the Life Sciences, the Eberly College of Sciences and the Institute of Cyberscience at Penn State, in part, under a grant with the Pennsylvania Department of Health using Tobacco Settlement Funds. The Department specifically disclaims responsibility for any analyses, interpretations, or conclusions. This research was also supported by the Czech Science Foundation (grant 15-02891S to EK). MoC was supported by the National Institutes of Health (NIH)-PSU funded Computation, Bioinformatics and Statistics (CBIOS) Predoctoral Training Program (1T32GM102057-0A1).

LIST OF SUPPLEMENTARY MATERIALS:

Supplementary Materials include Materials and Methods, 5 Supplementary Notes, Supplementary Methods, 9 Supplementary Tables, and 18 Supplementary Figures. Additionally, there are 2 Extended Data Files.

FIGURE LEGENDS

Figure 1. Non-B DNA motifs and Inter-Pulse Duration (IPD) analysis pipeline. A Non-B DNA motif types: patterns, putative structures, and counts of non-overlapping 100-bp windows containing one (and only one) motif with IPD measurements on the reference or reverse complement strand. **B** During SMRT sequencing, IPDs are recorded for each nucleotide in each subread (each pass on the circular DNA template). Subreads are aligned against the genome, and an IPD value is computed averaging ≥3 subread IPDs for each genome coordinate. We form 100-bp windows around annotated non-B motifs, extract their IPD values, and pool windows containing motifs of the same type to produce a distribution of IPD curves over the motif and its flanks (represented via 5th, 25th, 50th, 75th and 95th quantiles along the 100 window positions). We also form a set of non-overlapping 100-bp windows free from any known non-B motif, and pool them to produce a "motif-free" distribution of IPD curves. Each motif type is then compared to motif-free windows through Interval-Wise Testing (IWT). G4: G-quadruplexes; G4+/G4-: G4 annotated on the reference/reverse-complement strand; DR: direct repeats.

Figure 2. Polymerization kinetics at non-B DNA. A-D IPD curve distributions in motif-containing (red) vs. motif-free (blue) 100-bp windows, on reference (top) and reverse complement (bottom) strands. Thick lines: medians; dark-shaded areas: 25th-75th quantiles; light-shaded areas: 5th-95th quantiles. Red/blue marks (below top and above bottom plots): positions with IPDs in motif-containing windows higher/lower than in motif-free windows (IWT-corrected p-values ≤ 0.05). Heatmaps (between top and bottom plots): sequencing depth of motif-containing relative to motif-free windows (in percentages, can be >100%) on reference (Depth ref) and reverse complement (Depth rev) strands, and percentage of

windows with the motif (Motif) at each position. **A** G-quadruplexes. **B-D** STRs with disease-linked repeat number variation. **E** IWT results for IPD curve distributions in motif-containing vs. motif-free windows (reference strand). Each row shows significance levels (-log of corrected p-values) along 100 window positions for one motif type. White: non-significant (corrected p-value > 0.05); red/blue: significant, with IPDs in motif-containing windows higher/lower than in motif-free windows. STRs are grouped according to putative structure. **F** Comparison between observed mean IPDs in motif-containing windows and predictions from a dinucleotide compositional regression fitted on motif-free windows (reference strand). Bonferroni-corrected t-test p-values for differences: ≤0.0001 '****', ≤0.001 '***', ≤0.01 '**', ≤0.05 '*'. Black: non-significant (corrected p-value > 0.05); red/blue: significant, with observed mean IPDs higher/lower than composition-based predictions. Boxplot whiskers: 5th and 95th quantiles of the differences.

Figure 3. Relationship between G-quadruplex thermostability and polymerization kinetics. For the ten most common G-quadruplex motif types (G1 through G10, in order), we measured circular dichroism (delta epsilon) and light absorption (melting temperature, T_m), and computed average IPD values for each of thousands of motif occurrences in the genome (Table S5). For intramolecular G4s, average IPDs were regressed on: **A** delta epsilon (R-squared = 32.3%), and **B** T_m (5.7%; 9.2% if G9 is not considered). Boxplot whiskers: 5th and 95th quantiles. Boxplot width: proportional to the square root of the sample size for each motif. Points: individual occurrences used in the regressions, with horizontal jittering for visualization (results for intermolecular G4s are shown in Fig. S15).

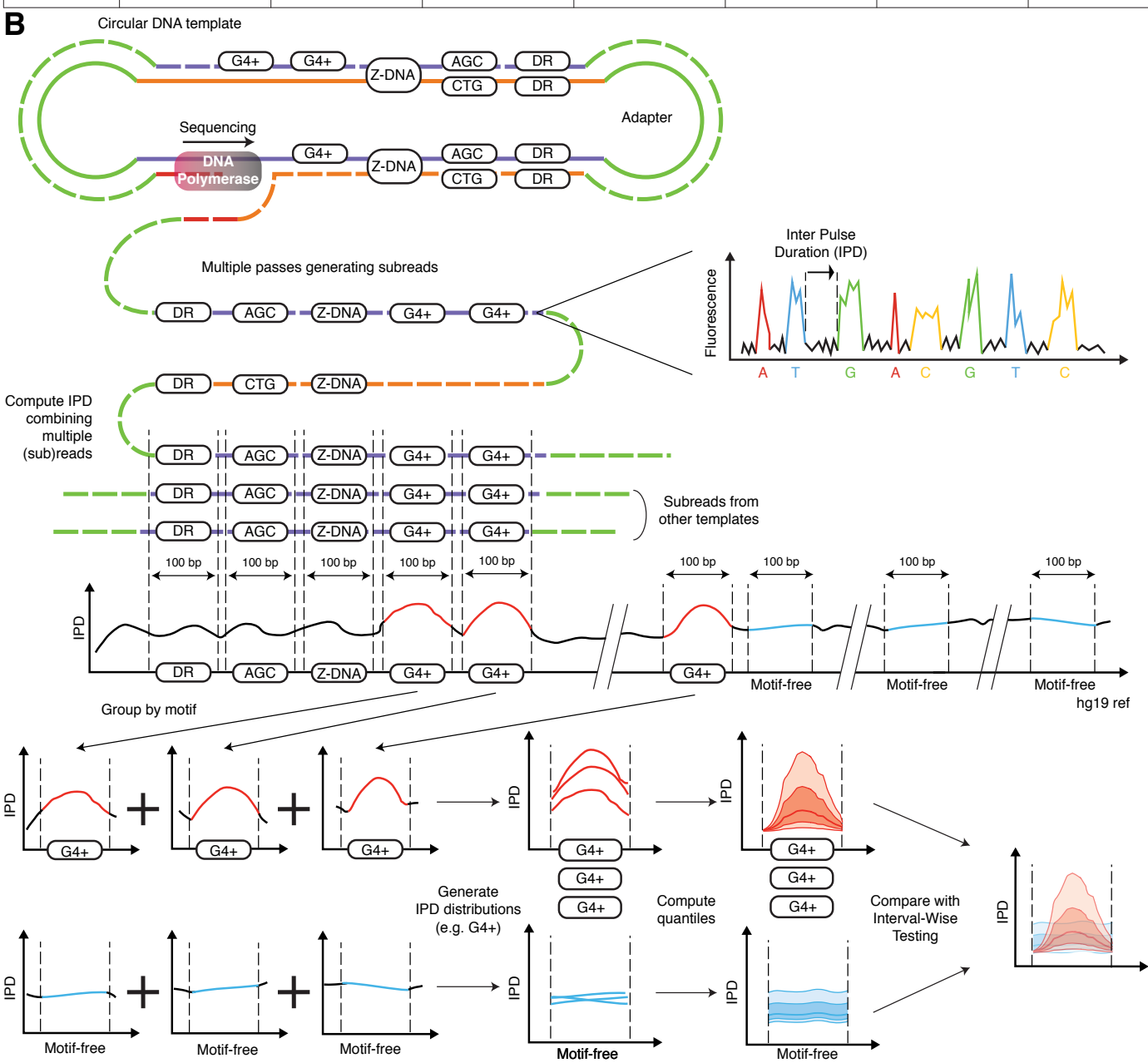
Figure 4. Effects of non-B DNA motifs on sequencing errors and mutations. For **A** mismatches, and **B** deletions, we contrasted rates of SMRT (PacBio) and Illumina sequencing errors, primate divergence (human-orangutan), human diversity (1,000

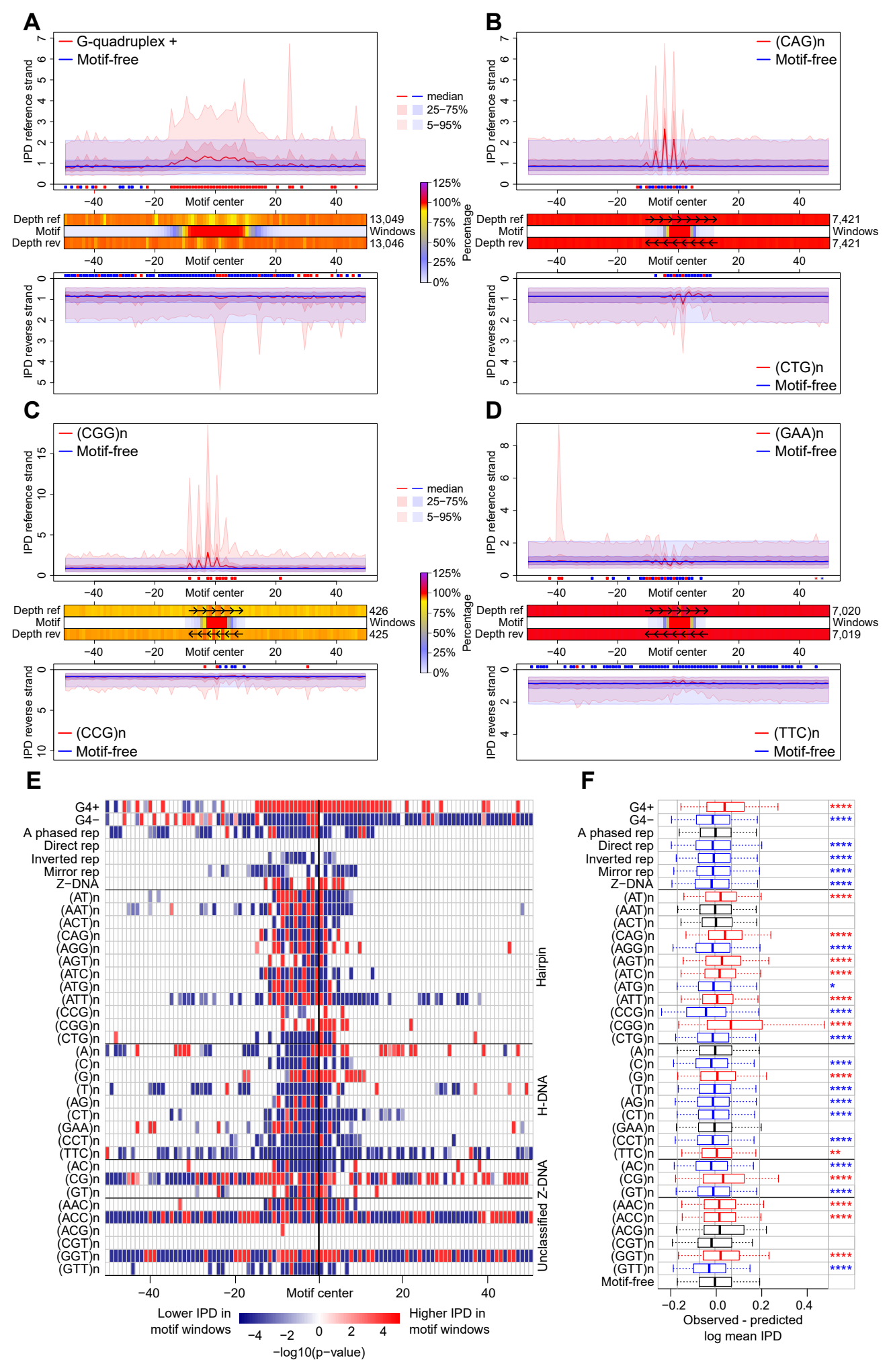
Genomes Project), and cancer somatic mutations (The Cancer Genome Atlas; TCGA) between motifs and motif-free regions. Results for insertions are shown in Fig. S17. White cells: insufficient events, inconclusive or non-significant (p-value > 0.10) results. Red/blue cells: rates in motifs significantly higher/lower than in motif-free regions. Numbers are fold differences (with proportional color intensities; 1-1.1 light, 1.1-2 medium, >2 dark) and test results (see Methods) are indicated by stars (p-value≤0.0001 '****', ≤0.001 '***', ≤0.01 '**', ≤0.05 '*', ≤0.10 '.'). Differences in significance are in part due to differences in sample sizes for the various datasets — PacBio having the largest (all sample sizes and details are reported in Tables S7-S8). For divergence as well as 1,000 Genomes and TCGA data, we cannot differentiate the strand where a mutation occurred, therefore a combined effect for both strands is shown. #SMRT mismatches fold-difference for Mirror Repeats is 1.003.

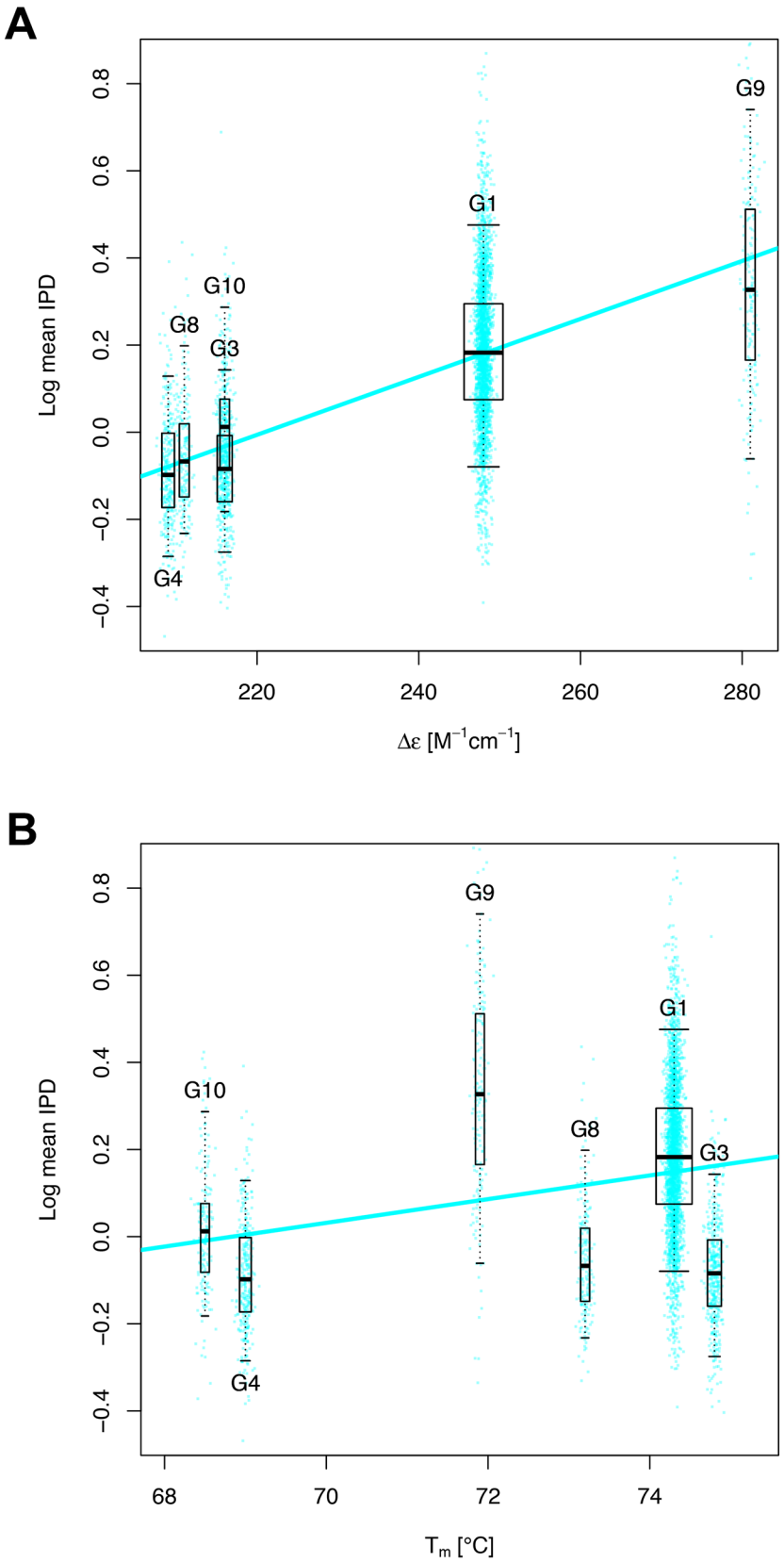
A

Motif	Inverted repeats	Mirror repeats	Direct repeats	Microsatellites	G4 motifs	Z-DNA motifs	A-phased repeats
Number of windows on ref strand	468,525	39,919	34,582	2,315,028	G4+ : 13,049 G4- : 12,876	6,207	26,142
Number of windows on rev comp	468,520	39,932	34,594	2,315,048	G4+: 13,046 G4-: 12,885	6,209	26,143
Motif pattern					GGG GGG GGG GGG	GYGYGYGYGYGYGYGYGY	AAAA AAAA AAAA 10 bp 10 bp
Structure	cruciform	H-DNA (triplex)	slipped strand	slipped strand	G quadruplex	Z-DNA	DNA bending

B




A
G-quadruplex +
Motif-free
IPD reference strand
IPD reverse strand
Motif center
Depth ref
Motif
Depth rev
13,049
Windows
13,046
median
25–75%
5–95%
Percentage
125%
100%
75%
50%
25%
0%
B
(CAG)n
Motif-free
7,421
7,421
(CTG)n
C
(CGG)n
426
425
(CCG)n
D
(GAA)n
7,020
7,019
(TTC)n
E
G4+
G4−
A phased rep
Direct rep
Inverted rep
Mirror rep
Z−DNA
(AT)n
(AAT)n
(ACT)n
(CAG)n
(AGG)n
(AGT)n
(ATC)n
(ATG)n
(ATT)n
(CCG)n
(CGG)n
(CTG)n
(A)n
(C)n
(G)n
(T)n
(AG)n
(CT)n
(GAA)n
(CCT)n
(TTC)n
(AC)n
(CG)n
(GT)n
(AAC)n
(ACC)n
(ACG)n
(CGT)n
(GGT)n
(GTT)n
Hairpin
H-DNA
Unclassified Z-DNA
Lower IPD in motif windows
Higher IPD in motif windows
−log10(p−value)
F
Motif-free
Observed - predicted log mean IPD


A
Log mean IPD
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
G9
G1
G10
G8
G3
G4
220
240
260
280
Δε [M⁻¹cm⁻¹]
B
Log mean IPD
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
G9
G1
G10
G8
G3
G4
68
70
72
74
Tm [°C]

A

Mismatches

	A-phased repeats	Direct repeats	Inverted repeats	Mirror repeats	Z-DNA	G4+ motifs	G4- motifs
SMRT		1.07 ****		1.00# **	1.19 ****	1.75 ****	
Illumina	1.20 **	4.29 ****	1.05 ***	1.16 **	1.53 *	2.71 ****	2.73 ****
Divergence	1.11 ****	1.59 ****	1.08 ****	1.02 **	1.78 ****	1.15 ****	
Diversity		1.03 ***	1.05 **		1.94 ****	1.30 ****	
TCGA	1.42 ***	2.27 ****	1.18 ****		2.55 ****		

B

Deletions

	A-phased repeats	Direct repeats	Inverted repeats	Mirror repeats	Z-DNA	G4+ motifs	G4- motifs
SMRT	1.04 ****	1.01 ****	1.01 ****		1.17 ****	1.46 ****	1.08 ****
Illumina		2.78 .				3.45 ****	4.77 ***
Divergence				1.20 .			
Diversity			1.39 *				
TCGA			1.39 ****	1.31 *		2.53 ****	