Title

Accurate sequencing of DNA motifs able to form alternative (non-B) structures.

Journal
Genome Research, 33(6)

Authors
Eckert, Kristin
Chiaromonte, Francesca
Huang, Yi-Fei
et al.

Publication Date
2023-06-01

DOI
10.1101/gr.277490.122

Peer reviewed

Accurate sequencing of DNA motifs able to form alternative (non-B) structures

Matthias H. Weissensteiner,[1,11,12] Marzia A. Cremona,[2,3,4,11] Wilfried M. Guiblet,[1,5] Nicholas Stoler,[6] Robert S. Harris,[1] Monika Cechova,[1,7,13] Kristin A. Eckert,[4,8] Francesca Chiaromonte,[4,9,10] Yi-Fei Huang,[1,4] and Kateryna D. Makova[1,4]

[1] Department of Biology, The Pennsylvania State University, University Park, Pennsylvania 16802, USA; [2] Department of Operations and Decision Systems, Université Laval, Quebec, Quebec G1V0A6, Canada; [3] Population Health and Optimal Health Practices, CHU de Québec–Université Laval Research Center, Québec, Quebec G1V4G2, Canada; [4] Center for Medical Genomics, The Pennsylvania State University, University Park, Pennsylvania 16802, USA; [5] Laboratory of Cell Biology, NCI-CCR, National Institutes of Health, Bethesda, Maryland 20892, USA; [6] Department of Biochemistry and Molecular Biology, The Pennsylvania State University, University Park, Pennsylvania 16802, USA; [7] Faculty of Informatics, Masaryk University, 60200 Brno, Czech Republic; [8] Department of Pathology, The Pennsylvania State University, College of Medicine, Hershey, Pennsylvania 17033, USA; [9] Department of Statistics, The Pennsylvania State University, University Park, Pennsylvania 16802, USA; [10] Institute of Economics and L'EMbeDS, Sant'Anna School of Advanced Studies, Pisa 56127, Italy

Approximately 13% of the human genome at certain motifs have the potential to form noncanonical (non-B) DNA structures (e.g., G-quadruplexes, cruciforms, and Z-DNA), which regulate many cellular processes but also affect the activity of polymerases and helicases. Because sequencing technologies use these enzymes, they might possess increased errors at non-B structures. To evaluate this, we analyzed error rates, read depth, and base quality of Illumina, Pacific Biosciences (PacBio) HiFi, and Oxford Nanopore Technologies (ONT) sequencing at non-B motifs. All technologies showed altered sequencing success for most non-B motif types, although this could be owing to several factors, including structure formation, biased GC content, and the presence of homopolymers. Single-nucleotide mismatch errors had low biases in HiFi and ONT for all non-B motif types but were increased for G-quadruplexes and Z-DNA in all three technologies. Deletion errors were increased for all non-B types but Z-DNA in Illumina and HiFi, as well as only for G-quadruplexes in ONT. Insertion errors for non-B motifs were highly, moderately, and slightly elevated in Illumina, HiFi, and ONT, respectively. Additionally, we developed a probabilistic approach to determine the number of false positives at non-B motifs depending on sample size and variant frequency, and applied it to publicly available data sets (1000 Genomes, Simons Genome Diversity Project, and gnomAD). We conclude that elevated sequencing errors at non-B DNA motifs should be considered in low-read-depth studies (single-cell, ancient DNA, and pooled-sample population sequencing) and in scoring rare variants. Combining technologies should maximize sequencing accuracy in future studies of non-B DNA.

[Supplemental material is available for this article.]

DNA conformations that deviate from the canonical right-handed double-helix with 10 nucleotides per turn are collectively termed "non-B DNA" (Zhao et al. 2010). The ability of the DNA molecule to fold into such alternative structures depends on the presence of certain sequence motifs (thereby called "non-B motifs"), which range in size from tens to hundreds of base pairs and account for a substantial portion of an organism's genome (e.g., ∼13% of the human genome) (Guiblet et al. 2018). Non-B DNA motifs can form distinct non-B DNA structures depending on their sequence (Fig. 1). A-phased repeat motifs, which consist of tracts of three to nine adenines or thymines (A-tract) separated by at least 4 bp (spacer), can facilitate bent double-helix structures (Koo et al. 1986; Barbič et al. 2003). In G-quadruplex (G4) motifs, which consist of at least four blocks of at least three guanines separated by one to seven arbitrary bases, the guanines from different blocks can bind to each other via Hoogsteen hydrogen bonds forming stems, with the arbitrary bases forming loops (Sen and Gilbert 1988; Burge et al. 2006). Direct repeat motifs, which consist of two copies of the repeated unit separated by a nonrepetitive spacer, can misalign, leading to slipped-strand structures with looped out bases (Sinden et al. 2007). Inverted repeat motifs, which consist of repetitive sequences complementary to each other (e.g., 5′-GACTGC and GCAGTC-3′) separated by a nonrepetitive spacer, are capable of forming hairpins and cruciform structures (Nag and Petes 1991). Mirror repeat motifs that consist of stretches of homopurines:homopyrimidines arranged in a mirrored fashion, separated by a spacer, can form triple-helix (H-DNA) structures (Htun and Dahlberg 1988). Finally, Z-DNA motifs, which consist of alternating pyrimidines and purines, such as $(CG:CG)_n$ or $(CA:TG)_n$, can

[11] These authors contributed equally to this work.
Present addresses: [12] Institute of Avian Research, D-26386 Wilhelmshaven, Germany; [13] Department of Biomolecular Engineering, University of California Santa Cruz, Santa Cruz, CA 95064, USA
Corresponding author: kdm16@psu.edu

Article published online before print. Article, supplemental material, and publication date are at https://www.genome.org/cgi/doi/10.1101/gr.277490.122. Freely available online through the *Genome Research* Open Access option.

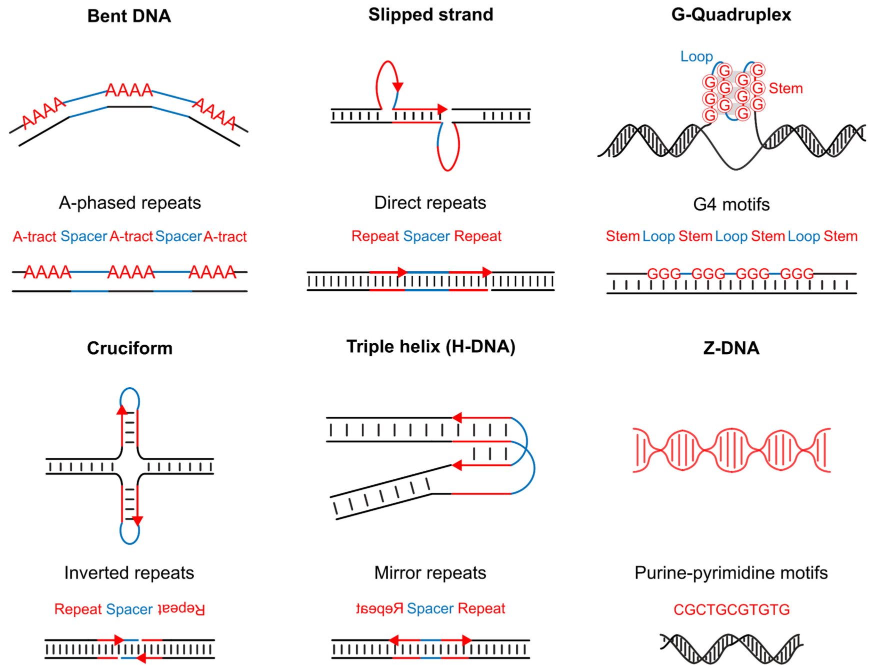

Figure 1. Types of non-B DNA structures. Shown are six types of alternative DNA structures (non-B DNA) with their respective underlying motifs and sequence arrangements. Mirror repeat motifs that form triple helix (H-DNA) structures consist of strongly skewed stretches of homopurines:homopyrimidines arranged in a mirrored fashion and separated by a spacer.

form left-handed zig-zag DNA structures (Wang et al. 1979; Singleton et al. 1982).

Non-B DNA structures can form in vivo (Biffi et al. 2013; Hänsel-Hertsch et al. 2016; Shin et al. 2016) and play an important role in essential cellular processes such as gene expression and DNA replication (Ghosh and Bansal 2003; Jain et al. 2008; Wang and Vasquez 2014). Non-B DNA may influence DNA synthesis because some non-B DNA structures have been shown to obstruct the progression and affect the accuracy of DNA polymerases (Mirkin and Mirkin 2007; Wang and Vasquez 2014). For instance, hairpins, G4 structures, Z-DNA, and triple helices have been linked to the inhibition of polymerase activity in vitro, with evidence illustrating a direct effect of realized non-B DNA structures on polymerase errors (Hile and Eckert 2004; Mirkin and Mirkin 2007; Hile et al. 2012; Stein et al. 2022). Furthermore, certain non-B-forming motifs were found to influence polymerase kinetics and affect sequencing errors of Pacific Biosciences (PacBio) sequencing (Guiblet et al. 2018).

Two major high-throughput sequencing technologies—Illumina and PacBio—are based on synthesis with DNA polymerases. In Illumina sequencing, the *Pyrococcus*-derived Phusion polymerase (Quail et al. 2012) is involved in the bridge amplification of the template strand, producing clusters, which is followed by sequencing by synthesis via the incorporation of fluorescently active nucleotides (Metzker 2010). In PacBio sequencing, an engineered bacteriophage phi29 DNA polymerase incorporates fluorescently labeled nucleotides (Eid et al. 2009), whose sequence is then determined by a laser (Logsdon et al. 2020). In contrast, in Oxford Nanopore Technologies (ONT) sequencing, DNA polymerase synthesis is absent, and the nucleotide sequence is determined by changes in electric current caused by the passage of a single-stranded DNA through a protein nanopore located in a synthetic membrane (Logsdon et al. 2020; Jain et al. 2016). Although it is currently unknown what effect non-B DNA structures might have on the activity of the engineered T4 phage Dda helicase, the motor protein used to unwind and propel the double-stranded DNA molecule during ONT sequencing (Daniel and Deamer 2019; Logsdon et al. 2020), there are examples of helicases being involved in the resolution of non-B structures (e.g., Jain et al. 2010). Although the conceptual approach of nucleotide sequence determination is considerably different among these three sequencing technologies, it is conceivable that the enzymes they recruit—polymerases and helicases—contribute to the technology-specific sequence error profiles. Because non-B DNA structures have been shown to affect DNA processing enzyme activity (Mirkin and Mirkin 2007), their effect on sequencing error profiles should be considered.

Errors have been a major concern ever since the invention of DNA sequencing, because they may drastically influence the downstream analysis and interpretation of sequencing data. Randomly occurring sequencing errors can be alleviated by increasing the read depth per site, enabling a consensus approach to identify the true nucleotide at a given locus (Nielsen et al. 2011). In cases in which the amount and/or quality of input DNA (e.g., ancient DNA studies) (Slatkin and Racimo 2016) or budget constraints do not allow high read depth, sequencing errors may have a major effect on the downstream analyses (Shafer et al. 2017). Sequencing errors that occur nonrandomly are expected to have an even greater impact, because such errors are expected to occur even with high read depth, resulting in them being more likely to be identified as false-positive genetic variants. Examples hereby are the coverage bias against GC-rich sequences in Illumina sequencing (Aird et al. 2011; Shafer et al. 2017) or the tendency of (earlier) versions of ONT sequencing to erroneously collapse homopolymer runs. Thus far, approaches to mitigate these issues have mostly included stringent computational filtering or the use of multiple independent sequencing technologies, which may drastically reduce the amount of usable data or may be cost-prohibitive, respectively.

In this study, we aimed at investigating a potential association between non-B DNA motifs and sequencing success for three major sequencing technologies (Illumina, HiFi mode of PacBio, and ONT). We compiled annotations of non-B-forming sequences in the human genome and used sequencing data from the Genome in a Bottle (GIAB) consortium (Zook et al. 2016) to detect errors. We used hypothesis testing, regression models, and probabilistic estimation to assess whether sequencing success is altered at non-B DNA and to differentiate between different factors contributing to technology-specific sequencing error profiles and false-positive variants in non-B motifs.

Results

We contrasted sequencing success, as measured by error rate, sequencing depth, and base quality, between motifs with non-B

DNA-forming potential ("non-B motifs") and control B DNA sequences. This was performed for sequencing reads generated with Illumina, PacBio, and ONT sequencing technologies. We used a data set in which these three technologies were applied to the same sample, an Ashkenazim son (HG002) from the GIAB consortium (Zook et al. 2016). For PacBio, the effects of non-B motifs on continuous long-read sequencing were evaluated previously (Guiblet et al. 2018), and thus, we focused on the HiFi circular consensus reads, which achieve low error rates after multiple passes over the same template. We acquired the genomic coordinates of A-phased repeat, direct repeat, inverted repeat, mirror repeat, and Z-DNA motifs from the non-B DNA database (Cer et al. 2013). The motifs potentially forming G4 structures were annotated with Quadron (Sahakyan et al. 2017).

Before assessing sequencing success, we applied two filtering schemes that differed in stringency. In the moderate filtering scheme, we only removed motifs with an average mappability score of less than one (see Methods), and we acknowledge that including the repetitive portions of the genome that are enriched in non-B motifs might introduce misalignments and ambiguous mapping of sequencing reads. With stringent filtering, we obtained "cleaner" (largely void of artifacts), but smaller, sets of non-B motifs by filtering out repeats and microsatellites, as well as overlapping motifs of different types. In the moderately filtered set, we restricted the size of motifs to the range from 10 to 1000 bp (mean = 12.33–58.25 bp, median = 11–51 bp) (Supplemental Table S1), excluded motifs that overlapped with other non-B motifs of the same type, and excluded motifs with an average mappability below one (see Methods). As a result, we retained 5,360,356 non-B motifs, covering ∼137 Mb of the genome (Supplemental Table S1). In the stringently filtered sets, we additionally removed non-B DNA motifs and controls that had ≥1-bp overlap with a repetitive element or a microsatellite, were within 50 bp from another non-B motif of any type, or had an average base quality of a Phred score below 30 for Illumina or below 73.2 for HiFi (these are corresponding thresholds for the two technologies; see Methods). As a result, we retained 710,553 motifs (13.5 Mb), 570,217 motifs (10.9 Mb), and 721,479 motifs (13.9 Mb) in the Illumina, HiFi, and ONT data sets, respectively (Supplemental Table S1). For each filtering scheme, each sequencing technology, and each type of non-B motif, we also generated a set of random control sequences, matching the corresponding motif set in number and length and excluding all non-B motifs, sequencing gaps, and ≥7-bp homopolymer runs. Before scoring sequencing errors, we removed biological variants as annotated in the GIAB true-variant set (Zook et al. 2016). We calculated two error rates for each type of mismatch error: per-motif rate and aggregate rate. The per-motif rate was obtained by dividing the total number of mismatch errors by the total number of aligned nucleotides for each motif and (separately) control sequence and then averaging them. The aggregate rate was calculated by summing up all mismatch errors and dividing by the total number of aligned nucleotides overall for motifs and control sequences.

Single-nucleotide mismatch error rates

Illumina

We detected significantly higher per-motif single-nucleotide mismatch (SNM) error rates for direct repeats, G4 motifs, and Z-DNA motifs compared with the respective controls for Illumina (for average per-motif and per-control error rates, see Fig. 2A, black circles; see also Supplemental Table S2A; for Benjamini–Hochberg-corrected *t*-test *P*-values, see Supplemental Table S4). This was the case for both moderately (1.35-, 2.39-, and 1.64-fold, respectively) and stringently (1.13-, 2.00-, and 1.49-fold, respectively) filtered sets (Fig. 2B). Error rates for A-phased and inverted repeats were significantly lower compared with the respective controls (0.90- and 0.93-fold for moderate filtering and 0.90- and 0.92-fold for stringent filtering, respectively), whereas mirror repeats had error rates significantly higher (1.20-fold) than that of the controls for the moderately filtered set and not significantly different from that of the controls for the stringently filtered set. Here and below, we also computed the aggregate error rates in motifs and corresponding controls (Fig. 2A, orange triangles; Supplemental Table S2A; Supplemental Fig. S1), which in most cases were similar to the average per-motif and per-control error rates. To illustrate that differences in error rates between motifs and controls are not solely because of a few outliers, we report the numbers and proportions of motifs and controls with at least one error of each error type in Supplemental Table S3.

To disentangle the factors contributing to the differences in average per-motif and per-control SNM error rates, we fit a Poisson regression model for each non-B motif type separately. In each model, the number of SNMs (in a motif or a control sequence) is the response, and the predictors are an indicator of "non-B motif versus control," the nucleotide composition, the motif length, the occurrence of 3- to 7-bp homopolymer runs, and the total number of sequenced nucleotides across all reads mapping to a motif (for details, see Methods). The results were similar between the moderately and stringently filtered sets (Supplemental Table S5); below, we discuss results for the former. To evaluate model performance and explanatory power, we calculated the percentage of deviance explained, that is, Cohen's pseudo-R^2, which ranged from 1.7% (for A-phased repeats) to 16.7% (for G4 motifs) (Supplemental Table S5). To estimate the contribution of each predictor, we followed the approach described by Kelkar et al. (2011) and fitted reduced models in which we left out individual predictors one at a time and calculated the reduction in deviance explained compared with the full model. Note, however, that, because of potential interactions between predictors, effects might not be additive. For the models with the highest deviance explained (G4 motifs and mirror repeats, 16.7% and 8.7%, respectively), the removal of the "non-B motif versus control" predictor reduced the explanatory power by 5.3% (G4 motifs) and 5.8% (mirror repeats). Likewise, removing nucleotide composition from the model substantially reduced the percentage of deviance explained in the models for G4 motifs (27.8%) and mirror repeats (by 97.8%). Here and below, such percentages among models should be compared by taking into account the overall deviance explained by each model (Supplemental Table S5).

To add an orthogonal approach to evaluate sequencing errors, we calculated SNM error rates as mismatches in the overlaps between Illumina read pairs in non-B motifs of the moderately filtered set, using sequencing data of the same individual as above (HG002) and following the method of Stoler and Nekrutenko (2021). Given that this approach identifies far fewer errors because it includes only the overlapping part of the reads, the resulting data set was admittedly smaller than the original one: In total, we analyzed between 494,961 (Z-DNA motifs) and 1,395,058 (inverted repeats) of overlapping nucleotides in non-B motifs (Supplemental Table S6). Overall, this analysis showed similar trends (Supplemental Table S6) to the analysis presented above. In particular, G4 and Z-DNA motifs showed substantially higher SNM error rates than did the controls, with 3.60-fold and 1.90-fold increases,

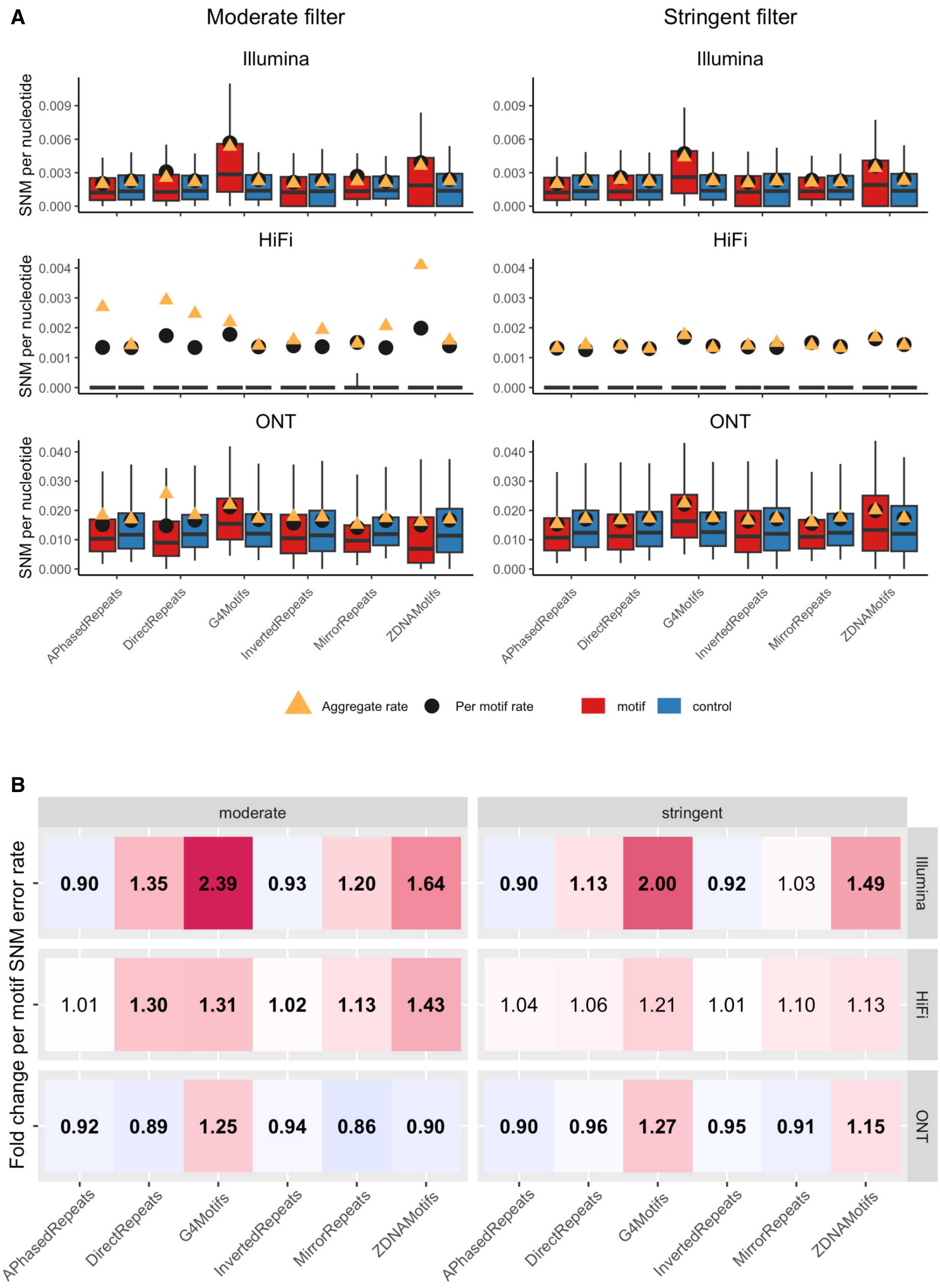

Figure 2. Single-nucleotide mismatch (SNM) error rates in non-B motifs. (*A*) Boxplots of per-motif SNM error rates. Boxplot whiskers show the fifth and the 90th percentiles, and values outside whiskers are excluded from the boxplots in order to better visualize the bulk of the distributions. The *left* panel shows the moderately filtered motif set; the *right* panel, the stringently filtered set. The three rows correspond to the different technologies (Illumina, HiFi, and ONT). Red and blue boxes correspond to motifs and controls, respectively; black dots mark per-motif means; and orange triangles aggregate error rates (sum of all errors divided by sum of all aligned nucleotides). Note that the *y*-axes differ among technologies. (*B*) Heat maps visualizing fold changes in per-motif means of SNM error rates between motifs and corresponding controls. Red (blue) shades indicate higher (lower) error rates in non-B motifs than in the controls, with fold change values reported in each cell of the map. When these values are in bold, per-motif means were significantly different between motifs and controls (*t*-test *P*-values corrected for multiple testing smaller or equal to 0.05). Also, here, *left* and *right* panels correspond to moderately and stringently filtered sets, respectively, and rows correspond to Illumina, HiFi, and ONT technologies, respectively.

respectively (Supplemental Fig. S2), whereas SNM error rates for the other non-B motif types were more similar to the controls and were even significantly decreased for direct and inverted repeats (chi-square test) (Supplemental Table S5).

PacBio HiFi

Even though SNM error rates for HiFi data were overall low (Supplemental Table S2B), they were significantly elevated for several non-B motif types compared with controls in the moderately filtered set (Fig. 2). Similar to Illumina, for HiFi we found significantly elevated per-motif SNM error rates compared with controls for direct repeats, G4 motifs, inverted repeats, mirror repeats, and Z-DNA motifs (1.30-, 1.31-, 1.02-, 1.13-, and 1.43-fold increases, respectively). The percentage of deviance explained for the Poisson regression models was usually low (Supplemental Table S5), and thus, the contribution of each predictor to the variation in SNM error rates could not be reliably determined. For the stringently filtered set, we found no significant differences between per-motif versus per-control SNM error rates (Fig. 2).

ONT

The overall ONT SNM error rate was an order of magnitude higher than that for Illumina or HiFi (Supplemental Table S2B). All comparisons of ONT SNM per-motif versus per-control rates, for both moderately and stringently filtered sets, were statistically significant (Fig. 2). Similar to Illumina and HiFi data, ONT reads showed higher per-motif SNM error rates in G4 motifs than in controls (1.25-fold and 1.27-fold for moderately and stringently filtered sets, respectively). However, fold differences of SNM error rates for the other non-B motifs versus controls were relatively small in magnitude (from 0.89-fold to 1.15-fold), and for Z-DNA, were inconsistent between the moderately and stringently filtered data sets. The explanatory power of the Poisson regression models for the ONT data was low (Supplemental Table S5); thus, the contribution of individual predictors could not be reliably determined.

SNM error rates for different parts of non-B motifs

We found a conspicuous pattern of variation in SNM error rates between different parts of non-B motifs, such as repeat arms and spacers in the repeat motifs and such as stems and loops in the G4 motifs (Fig. 3A; Supplemental Table S7). These patterns were qualitatively consistent between filtering schemes for all non-B DNA types and technologies (Fig. 3A). Below, we present fold differences for the moderately filtered set. Per-spacer error rates were significantly higher (Supplemental Table S7; for adjusted *P*-values, see Supplemental Table S8) than per-repeat-tract rates for A-phased repeats across all three technologies analyzed, with 1.40-fold for Illumina, 1.05-fold for HiFi, and 1.45-fold for ONT (Fig. 3B). Likewise, error rates were significantly elevated in spacers compared with repeat arms for direct repeats (1.39-fold for Illumina, 1.19-fold for HiFi, and 1.18-fold for ONT) (Fig. 3B). For inverted repeats, the rates were also significantly elevated in spacers compared with repeat arms in all three technologies, but the fold increases were small (≤1.10). For mirror repeats, error rates were significantly elevated in spacers compared with repeat arms for Illumina and ONT (1.23-fold and 1.16-fold, respectively) (Fig. 3B) but were the same between these two parts of repeats for HiFi. For G4 motifs, we observed contrasting patterns among technologies: Whereas error rates were significantly elevated in loops compared with stems in Illumina and HiFi (5.76- and 1.20-fold, respectively), they were decreased in loops versus stems (0.83-fold) in ONT (Fig. 3B). The pattern observed in ONT likely reflects the known elevated error rates at homopolymers (Bowden et al. 2019), which are present in G4 stems.

Deletion errors

Illumina

To compare deletion error rates between non-B motifs and controls, we divided the number of deletion errors by the number of aligned nucleotides (across all reads) of the motif or control sequence. The overall deletion error rates were low for the Illumina data set (Fig. 4A; Supplemental Table S2B). Deletion error rates were significantly higher for motifs than for controls in both stringently and moderately filtered sets for direct repeats (8.23- and 4.31-fold for the moderately and stringently filtered data sets, respectively), G4 motifs (3.00- and 3.02-fold), and inverted repeats (1.60- and 1.25-fold) (Fig. 4B; Supplemental Table S4). Additionally, deletion error rates were significantly higher in motifs than in controls in the moderately filtered set for Z-DNA motifs (12.7-fold), mirror repeats (4.84-fold), and A-phased repeats (1.20-fold) (Fig. 4B; Supplemental Table S2A). The percentage of deviance explained in the Poisson regression models for per-motif deletion error rates ranged from 0.4% (for A-phased repeats) to 16.3% (for Z-DNA motifs) (Supplemental Table S5). The removal of the "non-B motif versus control" predictor led to a reduction of deviance explained across all non-B motif types, with a 1.98%–83.2% reduction depending on the non-B motif type, with particularly high contribution of this predictor for models concerning Z-DNA motifs (83.2%) and direct repeats (64.5%) (Supplemental Table S5), which both also showed relatively high explanatory power for the full model (16.3% and 9.4%, respectively).

PacBio HiFi

The overall deletion error rate for HiFi was approximately an order of magnitude higher than that for Illumina (Supplemental Table S2B). All non-B motif types except for Z-DNA motifs showed significantly elevated per-motif deletion error rates compared with the controls (Fig. 4A), with 1.32-, 1.56-, 2.12-, 1.32-, and 1.67-fold increases for the moderately filtered set and 1.24-, 1.17-, 2.16-, 1.38-, and 1.58-fold increases for the stringently filtered set for A-phased repeats, direct repeats, G4s, inverted repeats, and mirror repeats, respectively (Fig. 4B; Supplemental Table S4). Per-motif deletion error rates in Z-DNA motifs were significantly reduced compared with the controls: 0.62-fold for the moderately filtered data and 0.24-fold for the stringently filtered data. The Poisson regression models for HiFi deletion error rates explained between 4.8% (Z-DNA motifs) and 12.0% (direct repeats) of deviance. Excluding the "non-B motif versus control" predictor led to a moderate reduction in percentage of deviance explained in models for direct repeats (17.5%) and mirror repeats (19.9%) (Supplemental Table S5). The presence of homopolymers was the most important predictor in all models; its removal led to the reduction of deviance explained between 57.9% and 91.7%.

ONT

The average per-motif deletion error rates were significantly different for all non-B motif types versus controls for the moderately filtered set and for G4s, inverted repeats, mirror repeats, and Z-DNA motifs for the stringently filtered set (Fig. 4A). Consistent elevation

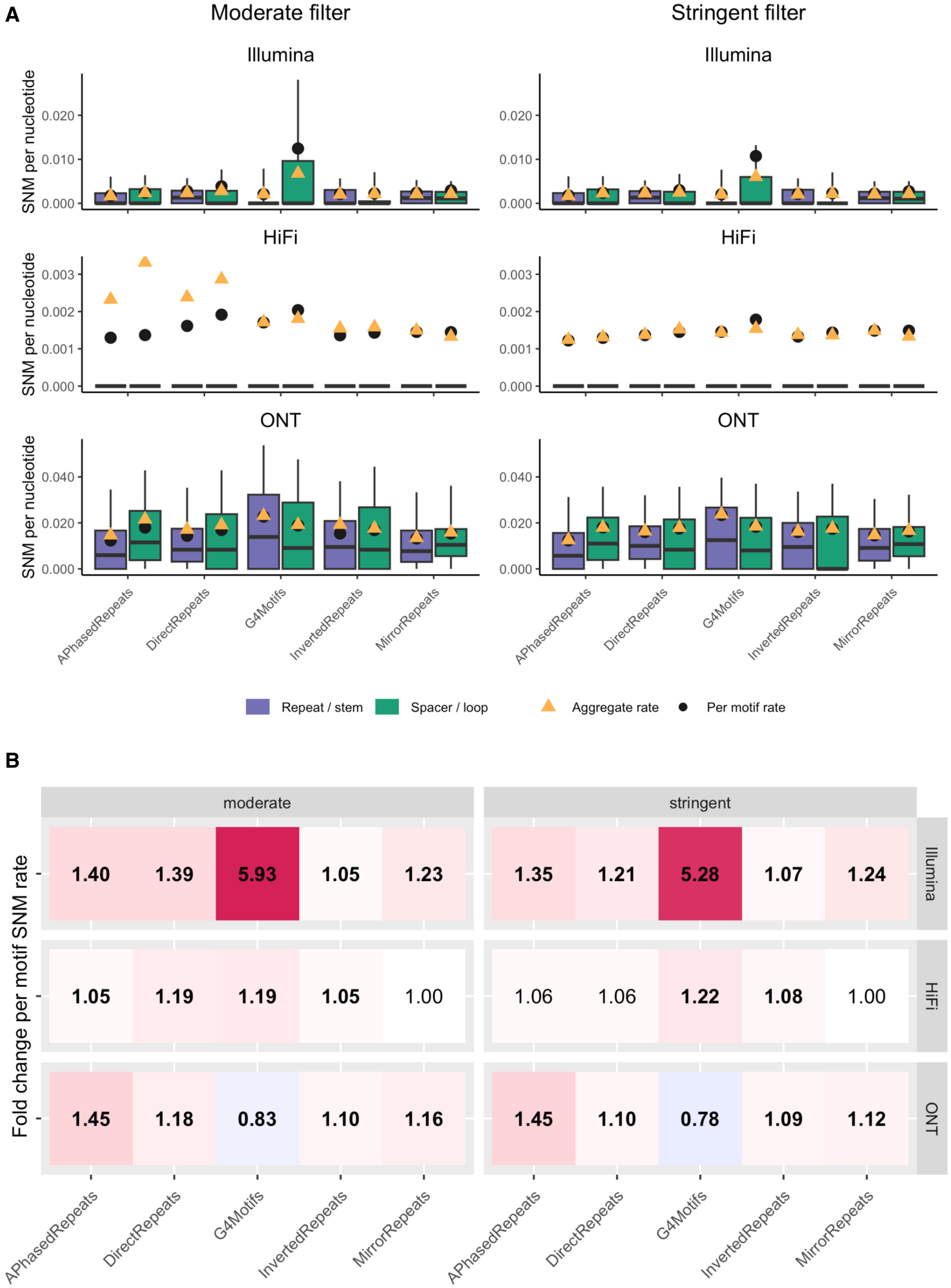

Figure 3. Single-nucleotide error rates in non-B motif subregions. (*A*) Boxplots of per-motif SNM error rates of subregions of non-B motifs. A-phased, direct, inverted, and mirror repeats are divided into repeat arms and spacer; G4 motifs are divided into stem (G-tract) and loop. Boxplot whiskers show the fifth and the 90th percentiles, and values outside whiskers are excluded from the boxplots in order to better visualize the bulk of the distributions. The *left* panel shows the moderately filtered set; the *right* panel, the stringently filtered set. The three rows correspond to the different technologies (Illumina, HiFi, and ONT). Purple and green boxes correspond to repeat/stem and spacer/loop subregions, respectively; black dots mark values for per-motif means; and orange triangles aggregate error rates (sum of all errors divided by sum of all aligned nucleotides). Note that the *y*-axes differ among technologies. (*B*) Heat maps visualizing fold changes in per-motif means of SNM error rates between different subregions of non-B motifs. Red (blue) shades indicate higher (lower) error rates in loops and spacers than in stems and repeat arms, with fold change values reported in each cell of the map. When these values are in bold, per-motif means were significantly different between motifs and controls (*t*-test *P*-values corrected for multiple testing smaller or equal to 0.05). Also here, *left* and *right* panels correspond to moderately and stringently filtered sets, respectively, and rows correspond to Illumina, HiFi, and ONT technologies, respectively.

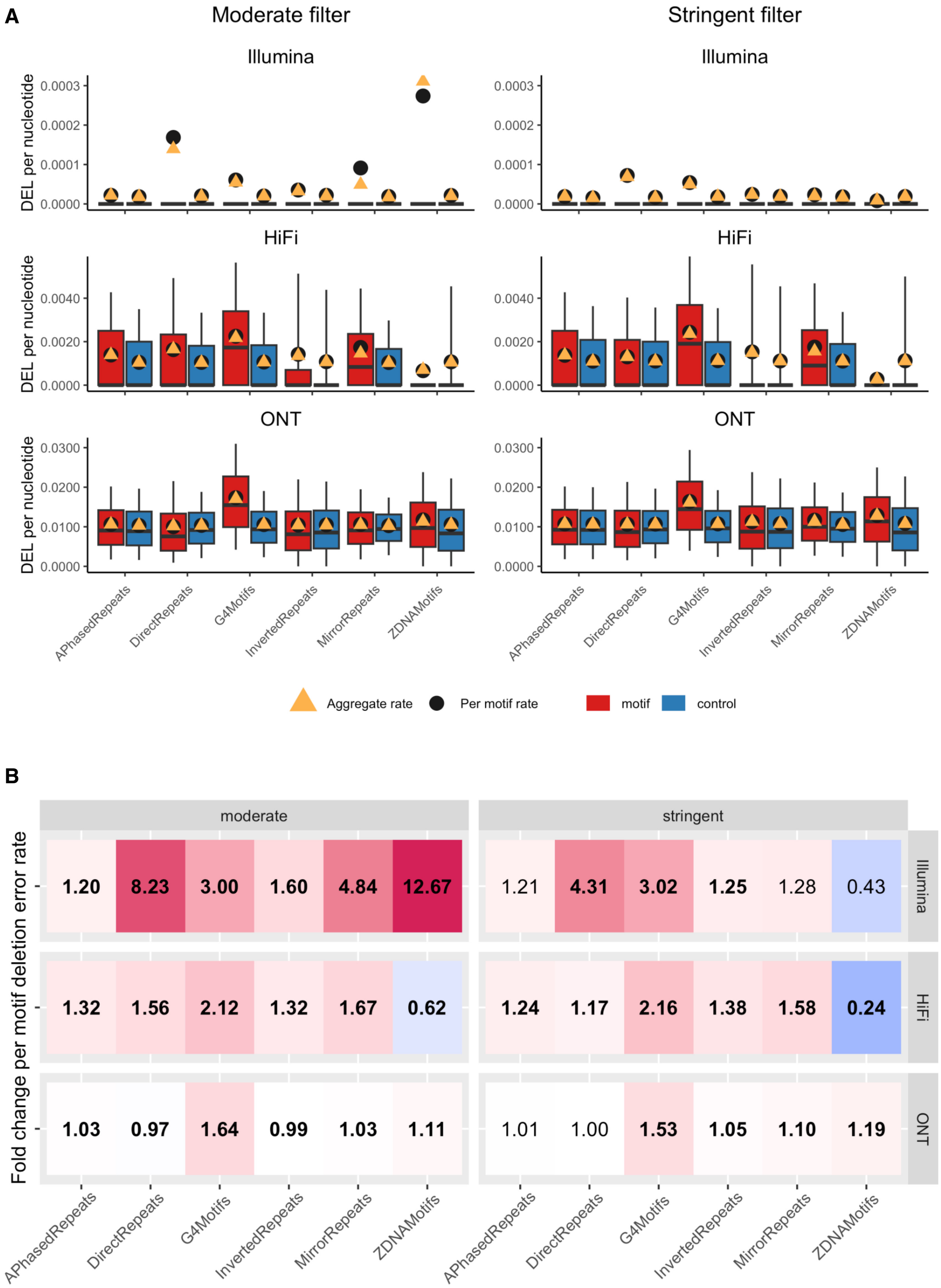

Figure 4. Deletion error rates in non-B motifs. (*A*) Boxplots of per-motif deletion error rates. Boxplot whiskers show the fifth and the 90th percentiles, and values outside whiskers are excluded from the boxplots in order to better visualize the bulk of the distributions. The *left* panel shows the moderately filtered motif set; the *right* panel, the stringently filtered set. The three rows correspond to the different technologies (Illumina, HiFi, and ONT). Red and blue boxes correspond to motifs and controls, respectively; black dots mark per-motif means; and orange triangles aggregate error rates (sum of all deletion errors divided by sum of all aligned nucleotides). Note that the *y*-axes differ among technologies. (*B*) Heat maps visualizing fold changes in per-motif means of deletion error rates between motifs and corresponding controls. Red (blue) shades indicate higher (lower) error rates in non-B motifs than in controls, with fold change values reported in each cell of the map. When these values are in bold, per-motif means were significantly different between motifs and controls (*t*-test *P*-values corrected for multiple testing smaller or equal to 0.05). Also here, *left* and *right* panels correspond to moderately and stringently filtered sets, respectively, and rows correspond to Illumina, HiFi, and ONT technologies, respectively.

in per-motif deletion rates was observed for G4 motifs over controls (1.64-fold and 1.53-fold for the moderately and stringently filtered sets, respectively) (Fig. 4B). Differences in deletion rates between motifs and controls were smaller in magnitude for mirror repeats, Z-DNA motifs, and inverted repeats (from 0.99- to 1.19-fold) (Fig. 4B). The explanatory power of the Poisson regression models explaining ONT deletion error rates ranged from 8.4% (for Z-DNA motifs) to 23.7% (for G4 motifs). For the latter, the removal of the "non-B motif versus control" predictor led to a reduction of deviance explained by only 0.7%, whereas in the model for Z-DNA motifs, we found an 18.9% reduction. In G4 motifs, the removal of the predictor denoting the presence of a homopolymer led to the reduction in deviance explained by 12.1%.

Insertion errors

Illumina

For Illumina, the insertion error rate was low and comparable to the deletion error rate (Supplemental Table S2B). The per-motif insertion error rates were significantly elevated in all non-B motif types compared with the controls for the moderately filtered set and for A-phased repeats, direct repeats, and G4 motifs for the stringently filtered set (Fig. 5A; Supplemental Table S2A). A-phased repeats, direct repeats, and G4 motifs showed a 1.50-, 29.2-, and 8.82-fold increase in insertion error rates in the moderately filtered set and a 4.78-, 3.65-, and 7.15-fold increase in the stringently filtered set, respectively (Fig. 5B). In the Poisson regression models, the percentage of deviance explained ranged from 1.00% (for A-phased repeats) to 21.5% (for Z-DNA motifs). Three Poisson regression models exceeded an explanatory power of 10%, namely, for direct repeats (16.7%), mirror repeats (20.5%), and Z-DNA motifs (21.5%). For these three models, we found that the removal of the predictor for "non-B motif versus control" led to a reduction of the deviance explained by 47.7%, 39.5%, and 83.8%, respectively.

PacBio HiFi

Similar to the deletion error rate, the insertion error rate for HiFi was approximately an order of magnitude higher than that for Illumina (Supplemental Table S2B). Average per-motif insertion error rates for HiFi were significantly elevated in all non-B motif types compared with controls in the moderately filtered set and in all non-B motifs but direct repeats in the stringently filtered set (Fig. 5A): with 1.23-, 1.40-, 1.07-, 1.32-, and 1.70-fold increases for the moderately filtered set and 1.21-, 1.30-, 1.08-, 1.20-, and 1.38-fold increases in A-phased repeats, G4 motifs, inverted repeats, mirror repeats, and Z-DNA motifs, respectively (Fig. 5B; Supplemental Table S4). The explanatory power of the Poisson regression models did not exceed 5% in any of the models, rendering only limited information about the contribution of individual predictors (Supplemental Table S5).

ONT

The insertion error rate in the ONT data set was higher than for other technologies analyzed (Fig. 5A; Supplemental Table S2B). Similar to that for the other technologies, for ONT, we found significantly elevated per-motif insertion error rates in G4 motifs compared with controls (1.60- and 1.52-fold for the moderately and stringently filtered sets, respectively) (Fig. 5B). For the other non-B motifs, insertion error rates were similar in magnitude between motifs and controls (with fold changes ranging between 0.98 and 1.08) (Fig. 5B), albeit often significantly different because of the large number of insertion events considered. Poisson regression models with ONT insertion error data explained between 2.4% (for Z-DNA motifs) and 22.5% (for G4 motifs) (Supplemental Table S5) of deviance. For the latter, the removal of the "non-B motif versus control" predictor led to a 2.5% reduction in the percentage of deviance explained.

Sequencing depth and quality

For Illumina, we found lower average read depth in G4 and Z-DNA motifs than in controls (0.81- and 0.95-fold, respectively, in the moderately filtered set and 0.82- and 0.90-fold in the stringently filtered set), and *higher* average read depth in A-phased, inverted, and mirror repeats than in controls (1.05-, 1.04- and 1.05-fold, respectively, in the moderately filtered set and 1.04-, 1.03- and 1.04-fold in the stringently filtered set) (Supplemental Fig. S3; Supplemental Table S2A). Direct repeats did not affect average read depth. For HiFi, we observed only minimal differences in the aggregate read depth between non-B motifs and controls, ranging from 0.98- to 1.01-fold for the moderately filtered set and from 0.99 to 1.01-fold for the stringently filtered set (Supplemental Fig. S3). For ONT, we found no relevant differences in aggregate read depth between non-B motifs and controls (Supplemental Fig. S3; Supplemental Table S4).

To evaluate potential effects of non-B DNA on sequencing base quality, we used the moderately filtered set. For Illumina, the average base quality was lower in direct repeat, G4, and Z-DNA motifs with 0.99-, 0.97-, and 0.96-fold differences compared with controls, respectively. Average base quality in A-phased repeat and inverted repeat motifs was slightly elevated compared with that in controls (1.02- and 1.01-fold, respectively), whereas it was equivalent between mirror repeat motifs and controls (Supplemental Fig. S4). For HiFi, average base quality was reduced in comparison to controls in direct repeat (0.99-fold) and Z-DNA motifs (0.96-fold), whereas it was increased in A-phased repeat (1.08-fold) and G4 motifs (1.08-fold). In inverted and mirror repeat motifs, average HiFi base quality was not different between motifs and controls. We could not measure the potential effects of non-B motifs on ONT base quality because no read base quality values were available in the data used (see Methods).

False-positive single-nucleotide variants owing to errors in non-B motifs

To evaluate the probability of identifying sequencing errors as variants (i.e., of false-positive single-nucleotide variants [SNVs]) and to gauge the expected number of such false positives in motifs with non-B-forming potential, we developed a probabilistic model that takes into account several key parameters (see Methods). It incorporates the per-nucleotide error rates derived from the analyses presented above, as well as the number of haploid genomes, the average sequencing read depth per haploid genome, the minimum number of reads used to identify a variant, the minor variant frequency used to call an SNV, and the total number of base pairs covered by non-B motifs (either of a certain type or all together; for details, see Methods).

We applied this model to three hypothetical scenarios with varying "sample sizes," that is, numbers of haploid genomes—200, 2000, and 20,000 (corresponding to 100, 1000, and 10,000 individuals)—and average read depths per haploid genome, ranging from 3× to 30× (Fig. 6A). For each scenario, we estimated the expected number of false-positive SNVs using the probabilistic model with the SNM error rates we computed above for the moderately

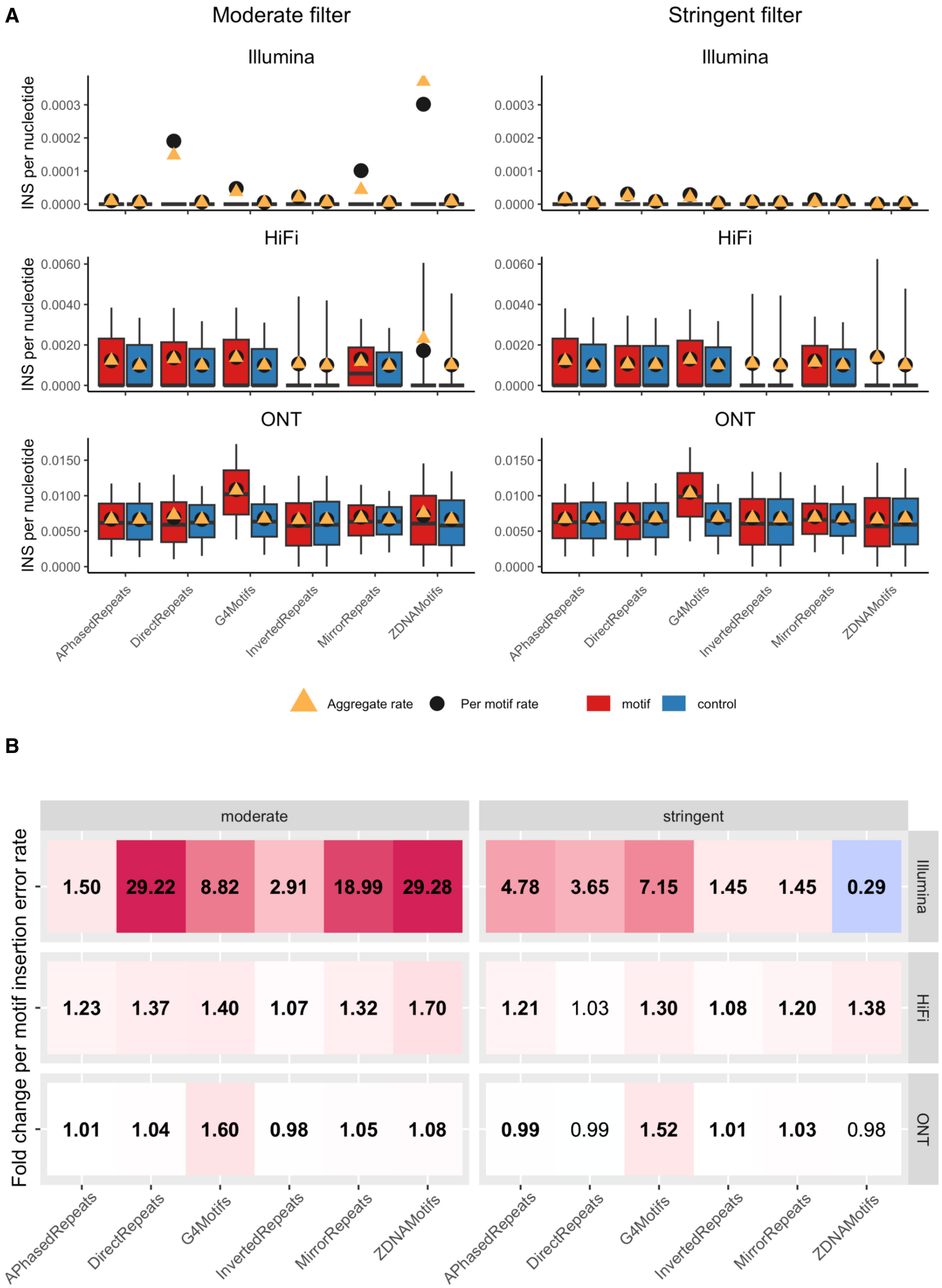

Figure 5. Insertion error rates in non-B motifs. (*A*) Boxplots of per-motif insertion error rates. Boxplot whiskers show the fifth and the 90th percentiles, and values outside whiskers are excluded from the boxplots in order to better visualize the bulk of the distributions. The *left* panel shows the moderately filtered motif set; the *right* panel, the stringently filtered set. The three rows correspond to the different technologies (Illumina, HiFi, and ONT). Red and blue boxes correspond to motifs and controls, respectively; black dots mark per-motif means; and orange triangles aggregate error rates (sum of all insertion errors divided by sum of all aligned nucleotides). Note that the *y*-axes differ among technologies. (*B*) Heat maps visualizing fold changes in per-motif means of insertion error rates between motifs and corresponding controls. Red (blue) shades indicate higher (lower) error rates in non-B motifs than in controls, with fold change values reported in each cell of the map. When these values are in bold, per-motif means were significantly different between motifs and controls (*t*-test *P*-values corrected for multiple testing smaller or equal to 0.05). Also here, *left* and *right* panels correspond to moderately and stringently filtered sets, respectively, and rows correspond to Illumina, HiFi, and ONT technologies, respectively.

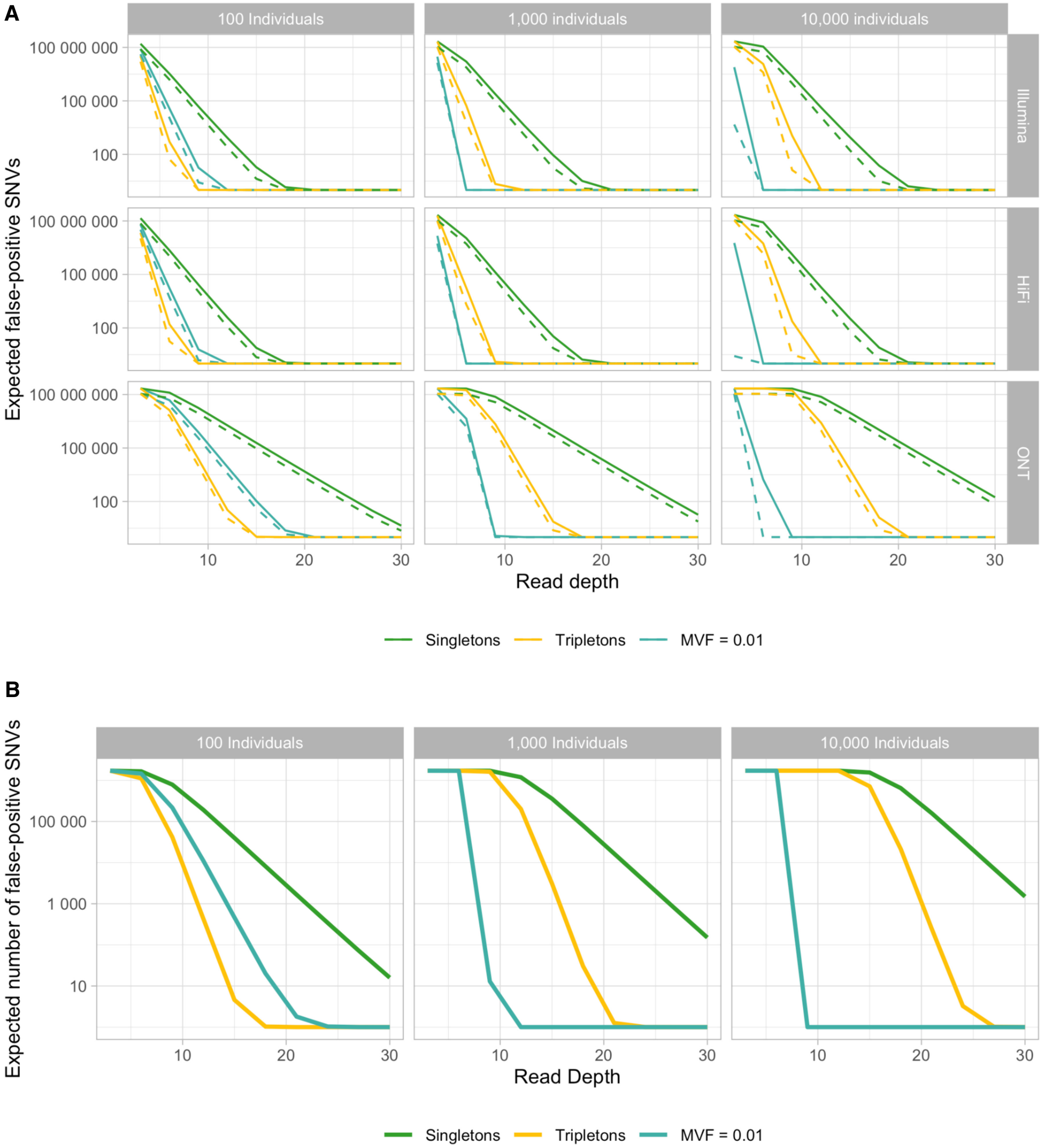

Figure 6. False-positive SNVs in non-B motifs. (*A*) Scenarios with different technologies corresponding to rows and numbers of diploid individuals (100, 1000, and 10,000) corresponding to columns. The average read depth per haploid genome is plotted on the *x*-axis, and the expected number of false-positive SNVs owing to errors is shown on the *y*-axis. Colors indicate different variant frequency filters (singletons, tripletons, and 1%) (for doubletons, see Supplemental Fig. S5), whereby the solid line corresponds to the cumulative number of all false-positive SNVs across non-B types; the dashed line, to the number of false-positive SNVs in an equally long stretch of B DNA. (*B*) Expected false-positive SNVs in middle guanines in guanine triplets at G4 motifs in Illumina sequencing. Within G4 motifs, there are 1,715,082 bp that fit the requirements of a middle guanine in a guanine triplet, which may have an extremely high error rate (Schirmer et al. 2016). Plotted are the numbers of expected false-positive SNVs, with the same coloring scheme and numbers of diploid individuals as in *A*.

filtered set in each of the three technologies. This was performed separately for motifs of each non-B type and for controls (B DNA) and considering different ways of scoring rare variants (i.e., by scoring variants present in a single haploid genome, by scoring variants present in three haploid genomes, and by using a minor variant frequency threshold of 0.01) (Fig. 6A; for results for doubletons, see Supplemental Fig. S5). Requiring a variant to occur in multiple (e.g., three to five) haploid genomes is frequently used when studying rare variants (Wainschtein et al. 2022). For all three ways of scoring variants, the expected number of false-positive SNVs was higher in all three technologies when considering non-B motifs of all types combined compared with an equally long stretch of B DNA sequence (solid and dashed lines, respectively, in Fig. 6A). However, above the read depth of 21× for Illumina and HiFi, the expected number of false positives approached zero even for singletons in all three sample size scenarios. In contrast, ONT sequencing may not be advisable when investigating singleton variants owing to its substantial expected false positives even at higher read depths. For tripletons and variants with a minor frequency above 0.01, expected false positives were much lower overall. Indeed, already at an average depth of 6×, virtually no false-positive SNVs are expected when requiring a minor frequency above 0.01 for sample sizes larger than 100 individuals in Illumina and HiFi (for ONT, this is achieved at approximately 9× read depth). Whereas motifs of all non-B DNA types were combined above, expected false positives computed separately for different non-B DNA types (Supplemental Fig. S6) depend, for each non-B motif type, on its error rate and on the number of nucleotides it occupies in the genome. In addition to calculating expected false positives, we also ran Monte Carlo simulations with our probabilistic model to gauge the variability of false-positive values under the different scenarios considered (see Methods), drawing congruent conclusions (Supplemental Fig. S7).

We then investigated a special case of an exceptionally high error rate potentially occurring in G4 motifs, namely, the middle guanines in guanine triplets. Schirmer et al. (2016) have shown that, in Illumina sequencing, the second positions of guanine triplets show error rates of orders of magnitude higher than genome-wide averages. Using this drastically elevated error rate (0.035 errors per site) in our probabilistic model for all potential middle G's in the G4 motifs analyzed in this study (a total of 1,715,082 bp), we obtained large expected false-positive values even at higher read depths (Fig. 6B) and at higher minor variant frequency cutoffs. Indeed, false positives for singletons did not approach zero even at sequencing depth of 30×, and they approached zero at the depth of 23× and 27× for tripletons scored for 1000 and 10,000 individuals, respectively. This indicates that for analyses of variants contained in G4 motifs, substantially higher read depths and/or more stringent minor variant filters are necessary to discern between true- and false-positive variants.

To further illustrate the potential impact of non-B motifs on false-positive SNVs, we used our probabilistic model with parameters derived from three publicly available sequencing data sets: the 1000 Genomes Project (1000G) (The 1000 Genomes Project Consortium 2015), the Simons Genome Diversity Project (SGDP) (Mallick et al. 2016), and the Genome Aggregation Database (gnomAD) (Karczewski et al. 2020). These data sets are all based on Illumina sequencing technology and have vastly different sample sizes as well as considerably different average sequencing read depths. In the 1000G example, which has an average read depth of 2× (1× per haploid genome) and a high number of individuals (5008 haploid genomes), singleton and tripleton variants cannot be reliably distinguished from sequencing errors, whereas requiring minor variant frequencies ≥0.01 drastically reduces the expected number of false positives (Supplemental Table S9). In contrast, in the SGDP example (21× read depth per haploid genome, 600 individuals), the sequencing depth is such that, according to our probabilistic model, false-positive variants are not expected regardless of the minor variant frequency. In the gnomAD example (15× read depth per haploid genome, 152,312 haploid genomes), 11,044 errors are expected to be falsely identified as singletons among all non-B motifs compared with 2481 errors among an equally long stretch of B DNA. These numbers are substantially reduced (leading to virtually no expected false positives according to our probabilistic model) in tripletons and variants with frequencies ≥0.01.

Finally, we also applied an empirical approach to the detection of false-positive SNVs. Randomly subsampling the Illumina data set used in the error detection analysis above to different read depths (3×, 9×, 15×, and 30×), we naively identified SNVs in the HG002 individual using FreeBayes (Garrison and Marth 2012) and then compared these variants to the GIAB true-variant set to estimate the proportion of false-positive SNVs (see Methods). In addition, we also used a newer version of the GIAB true-variant set (v4.2.1) to estimate false positives. Compared with the probabilistic model results (Fig. 6), with increasing depth, the proportion of false positives initially increased but then decreased (for both motifs and controls), starting at a read depth of 15× (Supplemental Fig. S8). The overall pattern of higher proportions of false positives in motifs compared with controls observed based on the GIAB true-variant set used in our main analysis (Supplemental Fig. S8A) was also evident when using a newer, more complete true-variant set GIAB v4.2.1 (Supplemental Fig. S8B). This suggests that true variants falsely identified as errors did not substantially contribute to such a pattern.

Discussion

Identifying biases in sequencing accuracy and predicting sequencing success are paramount for genomic studies. By using publicly available data, we showed that non-B DNA-forming motifs are associated with altered sequencing success across three major sequencing technologies. Thus, such motifs should be taken into consideration when interpreting existing sequencing studies and designing new ones. As our Poisson models suggested, the association of non-B motifs with error rates can be caused by the co-occurrence of other attributes, such as biased nucleotide composition or the presence of homopolymers. Yet, previous studies suggested that a variety of non-B DNA structures are affecting the function of DNA polymerases in vivo (Mirkin and Mirkin 2007), and this might be the case also when these enzymes are used in sequencing instruments.

SNM rates

Overall, we found moderate associations between non-B motifs and SNM error rates. The largest was observed for G4 motifs, which showed consistently elevated SNM error rates across all technologies and both filtering schemes (Fig. 2). The magnitude of this elevation was lower for HiFi and ONT than for Illumina (Fig. 2). For HiFi, when we restricted attention to the nonrepetitive portion of the genome through our “stringent” filtering, we observed no significant differences in SNM error rates between non-B motifs and controls (Fig. 2). For Illumina, our Poisson regression models suggest that in some cases altered SNM error rates are mainly associated with the presence of non-B DNA motifs themselves (e.g., Z-DNA, direct repeats, and G4 motifs); in others, with their peculiar nucleotide composition (e.g., mirror repeats and A-phased repeats). The latter is in line with previous studies indicating that GC content affects sequencing depth and errors in Illumina (Aird et al. 2011). We also found higher SNM error rates in spacers/loops than in repeat arms/stems, with particularly strong elevations for the Illumina technology and for G4 motifs (Fig. 3). This heterogeneity in SNM error rates within motifs should also be taken into account when analyzing variants located in non-B DNA. It suggests that although average error rates over an entire non-B motif might be only slightly elevated compared with control regions, certain subregions within the motif may be more prone to sequencing errors than others. This was also evident when we analyzed the expected number of false positives at the middle G's in G4s' stems (Fig. 6B).

Deletion error rates

The association of non-B motifs with deletion error rates was stronger than that with SNM error rates. Particularly for the Illumina technology, fold differences between non-B motifs and controls were larger in magnitude for deletion than for SNM error rates. G4 motifs had elevated deletion error rates over controls across technologies and filtering schemes, with the highest fold increases for Illumina, intermediate for HiFi, and lowest for ONT. Our models indicated that, in addition to the presence of non-B motifs, deletion error rates are strongly associated with the presence of homopolymers. Although ONT sequencing is known to show a bias in homopolymer regions (Bowden et al. 2019), the ongoing improvement of both base-calling algorithms and sequencing chemistry is expected to further reduce this bias in the future. Notably, although fold differences in deletion error rates between non-B motifs and controls were smaller in magnitude for HiFi than for Illumina and were smallest for ONT, both long-read technologies showed higher overall deletion error rates than the short-read Illumina.

Insertion error rates

We found that compared with deletion error rates, insertion error rates (albeit low) were more strongly affected by the presence of non-B motifs for Illumina, were less for HiFi, and were similarly largely unaffected for ONT. The fold increases in insertion error rates owing to the presence of non-B motifs for the Illumina technology were the largest across all the analyses we conducted. For this technology, even after applying stringent filtering, we found significant increases in insertion error rates. A notable exception are Z-DNA motifs, which seemingly become more accurate in the stringent filtering scheme. This is because of the removal of motifs overlapping with microsatellites, which appear to drive the high insertion, as well as deletion, error rate. Again for Illumina and for all non-B motif types but G4 motifs, our models indicated a contribution of the motif presence to explaining variability in insertion error rates. These contributions were only moderate for HiFi and minor for ONT.

Effects of moderate versus stringent filtering

We found that applying moderate versus stringent filtering alters Z-DNA deletion and insertion error rates for the Illumina technology. When the moderate filter is applied, Z-DNA shows highly elevated deletion and insertion error rates compared with that of controls, whereas these rates are reduced when the stringent filter is applied. Given that in the stringently filtered data set any overlap with microsatellites is removed, we suspect that microsatellites are driving the signal in the moderately filtered data set. We also note that our reanalysis of the SGDP data (Illumina sequencing) to compare diversity between B DNA and non-B DNA indicated G4 motifs stand out in terms of SNV diversity even when stringent filtering is applied (Supplemental Fig. S9). In addition to the two filtering levels, we have also explored the effects of using different true-variant sets in our error detection pipeline. For a subset of the whole data (Illumina and Chromosome 1), we repeated the error detection analysis using a more extensive true-variant set to identify errors and showed that although the overall SNV error rates were slightly reduced, the pattern of elevated rates in motifs compared with controls persisted (Supplemental Fig. S10).

Depth and quality

Across technologies, Illumina clearly showed the largest differences in average read depth between non-B motifs and controls and showed reduced read depths in G4 and Z-DNA motifs. G4 motifs showed the largest differences, with a 20% and 18% decrease in read depth compared with the controls for the moderately and stringently filtered sets, respectively. Read depth was hardly altered at all for the two long-read technologies, HiFi and ONT. For Illumina and HiFi technologies, the associations between non-B motifs and sequencing quality appear to be minor.

Conclusions and recommendations

Although our results suggest a relationship between non-B motifs and sequencing accuracy, providing evidence for a causal link between non-B DNA structure formation and sequencing errors is more difficult. Confounding factors, such as biased GC content, which are also known to influence sequencing accuracy in at least one of the technologies (Illumina), are in close interplay with non-B motifs, as they are often an inherent feature of such motifs (Dohm et al. 2008). The same is true for homopolymers (including imperfect homopolymers occurring in G4 motifs), which are known to elevate the error rates of both HiFi and ONT technology (Bowden et al. 2019; Karst et al. 2021). We conclude that many non-B motifs possess several attributes associated with elevated sequencing error rates (structure, biased nucleotide composition, and homopolymers), with effects and magnitudes differing across technologies and sequencing error types. Therefore, the choice of technology depends on the type of errors one is trying to avoid at non-B motifs. To minimize single-nucleotide errors at non-B motifs, we recommend using PacBio HiFi, particularly for the nonrepetitive portion of the genome. Both HiFi and ONT display low SNM error biases at non-B DNA motifs compared with Illumina; however, HiFi has lower overall SNM error rates compared with those of ONT. If minimizing deletion and insertion error biases at non-B motifs is of interest, the choice should be between HiFi and ONT, which show comparatively lower biases at non-B motifs, with a preference toward HiFi, which has overall low indel rates. Illumina has low insertion and deletion error rates, but its insertion and deletion error biases at non-B motifs are substantial, making it a suboptimal choice.

If one is to choose one sequencing technology to obtain the most accurate results for non-B motifs and minimize all three types of sequencing errors, we would recommend HiFi, which balances out low error rates with relatively low biases at non-B motifs. The ONT technology, although having higher overall rates for all errors considered, does not appear to be affected by non-B DNA motifs as much as Illumina and PacBio HiFi, indicating that ONT may carry less bias in non-B motifs. This is consistent with the fact that this technology is not polymerase based (Daniel and Deamer 2019). However, G4 motifs are the only type of non-B that shows a considerable increase of mismatch errors, which is somewhat surprising because the helicase used in ONT sequencers should, in principle, be less susceptible to structures formed in single-stranded DNA. Overall, given a sufficient read depth, ONT might be less prone to false-positive variants at non-B motifs, which, in addition to its ability to generate ultra-long sequencing reads (Jain et al. 2018), makes it attractive.

The use of multiple technologies facilitates error detection for any type of mutation, because true variants should be present in all technologies, each with its own sequencing error profile. In addition to combining different technologies, our calculation of expected false-positive SNVs based on SNM error rates in non-B motifs suggests that read depth and variant frequency cutoffs are critical, especially for rare variants. The high amount of expected false-positive singleton variants in scenarios with low read depth, and particularly in subregions of G4 motifs, highlights the importance of treating these regions with extra caution in downstream analyses and of applying rigorous quality filters. Moreover, results obtained from our probabilistic model with parameters derived from the 1000 Genomes, SGDP, and gnomAD data sets highlight how investigating rare variants with low read depth may be particularly problematic. In some extreme cases (e.g., middle guanines in guanine triplets) with insufficient read depth, variants may become indistinguishable from errors if they are only occurring once or a few times among studied individuals. Taken together, these results are in line with previous findings on the relationship between read depth, minor variant frequency, and the occurrence of false positives. They highlight the need for sufficient depth and quality control in error-prone regions (Tabangin et al. 2009; Kishikawa et al. 2019), to which we now add non-B DNA, and suggest that investigations incorporating rare variants in such regions should be performed with additional caution and sufficient read depth.

Overall, in cases in which abundant read depth is available, the increased error rate in non-B motifs might not lead to mistaking errors for biological variants (i.e., to false positives), especially when stringent filters are applied in terms of read quality and the avoidance of repetitive regions. However, when read depth is low (e.g., in ancient DNA or pooled population sample studies), variants identified in non-B motifs require additional scrutiny before being used in downstream analyses, and it may be advisable to restrict attention to variants identified by multiple technologies (when available) to avoid technology-specific biases.

Methods

Data

We used publicly available sequencing data generated for the GIAB Consortium (Zook et al. 2016; https://www.nist.gov/programs-projects/genome-bottle). We downloaded Illumina, PacBio HiFi, and ONT data for one individual—the son of the Ashkenazim trio (HG002, NA24385). We down-sampled the existing Illumina (2 × 150 bp, generated with Illumina HiSeq 2500 Rapid SBS) alignment file of 300× to ∼100×. For PacBio HiFi, we downloaded 167 Gb of raw read data (30× consensus read depth, generated with the PacBio Sequel instrument). Because the ONT base-calling is under constant development, we used a data set available at EPI2ME Labs (https://labs.epi2me.io), which uses a significantly improved base-calling algorithm (Bonito v0.3.0) on the same raw data set with sequencing depth of 57× (HG002, NA24385).

Read mapping

In all our analyses, we used hg19 as a reference, as the most comprehensive genomic annotations and other resources are available for this version. Because we are applying a variety of filters to remove repetitive regions (see below), we do not expect our analysis to be influenced by the choice of this reference. For Illumina, we downloaded alignment files from the GIAB homepage (https://github.com/genome-in-a-bottle/giab_data_indexes). For PacBio and ONT, we aligned reads to hg19 using minimap2 with the PacBio HiFi and ONT specific parameters, respectively (Li 2018). In all alignment files, we removed duplicates using Picard tools (http://broadinstitute.github.io/picard/) and sorted and split reads into forward and reverse and by chromosome using SAMtools (Li et al. 2009).

Non-B DNA annotations

For all non-B DNA motifs except G4 motifs, we used annotations available at the non-B DNA database (https://nonb-abcc.ncifcrf.gov), which are based on the human reference hg19. To predict potentially G4-forming loci, we used Quadron (Sahakyan et al. 2017), which provides predicted stability values for each motif, with default parameters. We then downloaded the mappability track for hg19 based on 36-mers from the UCSC Genome Browser (http://genome.ucsc.edu/), used the R package genomicRanges (Lee and Schatz 2012; Lawrence et al. 2013) to calculate mean mappability values, and used BEDTools nuc (Quinlan and Hall 2010) to obtain nucleotide composition (abundance of each of the four bases) for all motifs. For base quality, we first obtained values for each read and position within each motif using SAMtools mpileup (version 1.9) (Li et al. 2009) and then calculated the average for each motif. Because Illumina and PacBio technologies use different quality score encoding, these results are not directly comparable. For the ONT data set, there are no quality scores available in base-called read data.

For each non-B motif type, we removed overlapping motifs (within the same non-B type) and those overlapping with a nucleotide homopolymer ≥7 bp, as well as any motif >1000 bp or with an average mappability lower than one. In addition, we recorded any overlap with a motif of another non-B type, a RepeatMasker annotation, a homopolymer ≤7 bp, or an annotated microsatellite. Microsatellite annotations were generated with STR-FM (Fungtammasan et al. 2015), identifying mono-, di-, tri-, and tetranucleotide repeats with a copy number of at least seven units. This set of non-B motifs, which we call the “moderately filtered set,” formed the basis for randomly generating control regions: Independently for each non-B type, we constructed controls matching the number and the size of the motifs and excluding reference gaps and all non-B DNA motifs. Finding controls with matched nucleotide composition was not feasible for some non-B DNA motif types owing to their high number and extensive sequence coverage, especially when motifs had biased GC content (e.g., G4 motifs). For all control regions, we gathered features such as base quality, nucleotide composition, overlap with RepeatMasker annotations, etc. To form the “stringently filtered set,” we first excluded any motif that overlapped with a RepeatMasker or an STR annotation and any motif that either overlapped with or was within 50 bp of another non-B motif. Additionally, we excluded any motif with a Phred quality score lower or equal to 30 (Illumina) or 73.17 (HiFi). To adjust controls, we subjected them to the same filters and then randomly subsampled them to match the size of the corresponding motif sets.

To assess variation in error rate within non-B motifs, we further partitioned annotated motifs into subregions. For A-phased, direct, inverted, and mirror repeats, we split the annotations into repeat arms and spacers; for G4 motifs, into stems (the G-tract) and loops. To be consistent across motifs, we restricted analyses on subregions to G4 motifs with four stems and three loops and to repeat motifs (A-phased, direct, inverted, and mirror repeats) with two repeat arms and one spacer.

Error calling

Because variant detection tools are usually focused on detecting true biological variation and are optimized to avoid sequencing errors, we developed a script that, for each region of interest, counts the total number of aligned nucleotides and naively identifies mismatches directly from the CIGAR string of an alignment file. To exclude biological variants not representing sequencing errors from our analyses, we used the HG002 GIAB true-variant set previously described (Zook et al. 2016). Briefly, the variant set was obtained by mapping all sequencing data of the Ashkenazim Jew trio to the hg19 reference and then jointly calling variants using GATK HaplotypeCaller on all three samples. The resulting VCF file was filtered following GATK SNP variant quality score recalibration and GATK best practices recommendations (Zook et al. 2016). In our error detection pipeline, any single-nucleotide, insertion, or deletion mismatch was recorded as an error unless it overlapped with a true biological variant present in the true-variant set described above. In the latter case, it was removed from all subsequent analyses, and the position was treated as if no mismatch had occurred for HG002 (Zook et al. 2016).

To add an orthogonal approach in single-nucleotide error calling for Illumina, we also detected SNV sequencing errors by examining the overlaps between mates in Illumina read pairs as described by Stoler and Nekrutenko (2021), using the full (300×) data set described above. For each read pair, we restricted our analysis only to the errors located in the region between 50% and 60% of the full read length. For this analysis, we used the same sets of non-B motifs and controls as described above.

Read depth and base quality

To assess potential biases in sequence read depth, we calculated the mean depth per base pair (total number of aligned nucleotides divided by total length) and per motif (averaging the mean depth per bp across all motifs) for all non-B motif types and associated controls. Likewise, we computed mean base quality for motifs and controls for Illumina and HiFi sequencing data.

Downstream statistical analysis

All statistical analyses were performed separately on the "moderately" and "stringently" filtered sets described above. The *t*-test *P*-values for the comparisons between the average per-motif and per-control mismatch error rates were adjusted for multiple testing using the Benjamini–Hochberg correction (Benjamini and Hochberg 1995).

To study the effect of non-B-forming motifs on sequencing errors while taking into consideration the effects of other quantities such as nucleotide composition, motif length, and the presence of homopolymers, we fitted a Poisson regression model using the glm function in R with Poisson distribution and log link function. This model was chosen because errors, that is, our responses, displayed extremely right-skewed distributions with an excess of zero. Specifically, we used the counts of single-nucleotide mismatches, deletions, and insertions as responses in separate regressions, always including the logarithm of the total number of sequenced nucleotides as an offset term, namely, a predictor with fixed coefficient equal to one, to control for motif length and sequencing depth. In symbols, we used the Poisson model

$$\log(E(\#err \mid x)) = \log(\#\,nucl) + \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p,$$

where $\#\,err$ is the error count, $\#\,nucl$ is the total number of sequenced nucleotides, and x_1, …, x_p are the predictors. This model is mathematically equivalent to using the error rates as responses:

$$\log\left(\frac{E(\#\,err \mid x)}{\#\,nucl}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p.$$

Nucleotide composition, as represented by the proportions of A, C, T, and G nucleotides in each motif or control, was transformed using the isometric log-ratio transform of Aitchison for compositional data (Aitchison 1982). For each regression, after fitting the full model with all predictors included, we excluded influential outliers (i.e., observations with a Cook's distance greater than one) (Cook and Sanford 1982), fit the model again, performed a goodness-of-fit chi-square test, and computed Cohen's pseudo R^2; that is, the share of deviance explained (D_0–D_m)/D_0, wherein D_0 is the null deviance and D_m the residual deviance of the model. To estimate the contribution of each individual predictor, we then repeated the fit, excluding from the model one predictor at a time and calculating the reduction in the share of deviance explained (Kelkar et al. 2011); in symbols, $[(D_0 - D_m) - (D_0 - D_{m\{-\}})]/(D_0 - D_m)$, where D_m and $D_{m\{-\}}$ are the residual deviances of the full and reduced model, respectively. All statistical analyses were performed with the R programming language, and figures were produced with the ggplot2 package (Wickham 2011; R Core Team 2022).

Probabilistic model and empirical approach to estimate false-positive SNVs

We built a probabilistic model to quantify the contribution of sequencing errors in variant identification and to estimate the effect of non-B motifs on the number of false-positive SNVs called in several scenarios. Specifically, we modeled the presence of a sequencing error at a single site of a certain type (e.g., a site belonging to a certain type of non-B motif) using the Bernoulli distribution; that is, $X \sim B(1, r)$, where r is the corresponding per-nucleotide sequencing error rate. Assuming independence among errors of different sequenced nucleotides mapping to the same site, the number of errors observed in a site sequenced at depth d is $Y \sim B(d, r)$, a binomial distribution with d trials. Further assuming that all the sequencing errors in a site generate the same variant (worst-case scenario), the probability of wrongly identifying a variant in a site of a haploid genome is $p_{var} = P(Y \geq \min_{reads})$, where $\min_{reads}$ is the minimum number of reads required to call a variant. Finally, considering the total number of haploid genomes g and assuming independence among their sequencing errors, we obtain that the number of haploid genomes with a variant in a site owing to sequencing errors can be modeled as $V \sim B(g, p_{var})$. Hence, the probability of identifying a variant in a site owing to sequencing errors is $p_{SNV} = P(V \geq \min_{var})$, where $\min_{var}$ is the minor variant frequency to call an SNV. Using such probability p_{SNV} and the total number of sites of the considered type and assuming constant sequencing depth at all sites, we can compute the expected number of false-positive SNVs as $p_{SNV} \cdot n_{sites}$. Across the different read depth scenarios (3× to 30×, in increments of 3), $\min_{reads}$ was always one-third of the respective read depth. Because this calculation only provides an expected number (an average) of false-positive SNVs, we also implemented a Monte Carlo simulation study based on the same generative model to evaluate the corresponding variability in false-positive values under various scenarios.

In addition, we estimated the proportion of false-positive SNVs using the same Illumina data set that was also used for the error detection analyses. To illustrate the effect of varying read depths on false positives, we subsampled the HG002 100× data set to approximately 3×, 9×, 15×, and 30× and performed variant calling using the tool FreeBayes with default parameters and a minimum mapping quality filter set to 30 (Garrison and Marth 2012). We then intersected the obtained variant set with the combined annotation of non-B motifs and controls and compared these variants with the GIAB true-variant set described above (v3.3.2), as well as with a newer, more complete version (v4.2.1) (Wagner et al. 2022). Called variants present in the true-variant set were scored as true positives; those absent, as false positives. We then calculated the proportion of false positives by dividing the number of false positives by the sum of true and false positives for both motifs and controls.

Software availability

All custom scripts used in our analyses are available at GitHub (https://github.com/makovalab-psu/nonB-Seq-Errors) and as Supplemental Code.

Competing interest statement

The authors declare no competing interests.

Acknowledgments

This project was supported by National Institutes of Health grants R01GM136684 (to K.D.M.) and R01CA23715 (to K.A.E.).

References

The 1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. *Nature* **526:** 68–74. doi:10.1038/nature15393

Aird D, Ross MG, Chen W-S, Danielsson M, Fennell T, Russ C, Jaffe DB, Nusbaum C, Gnirke A. 2011. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol* **12:** R18. doi:10.1186/gb-2011-12-2-r18

Aitchison J. 1982. The statistical analysis of compositional data. *J R Statist Soc Ser B* **44:** 139–160. doi:10.1111/j.2517-6161.1982.tb01195.x

Barbič A, Zimmer DP, Crothers DM. 2003. Structural origins of adenine-tract bending. *Proc Natl Acad Sci* **100:** 2369–2373. doi:10.1073/pnas.0437877100

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Statist Soc Ser B* **57:** 289–300. doi:10.1111/j.2517-6161.1995.tb02031.x

Biffi G, Tannahill D, McCafferty J, Balasubramanian S. 2013. Quantitative visualization of DNA G-quadruplex structures in human cells. *Nat Chem* **5:** 182–186. doi:10.1038/nchem.1548

Bowden R, Davies RW, Heger A, Pagnamenta AT, de Cesare M, Oikkonen LE, Parkes D, Freeman C, Dhalla F, Patel SY, et al. 2019. Sequencing of human genomes with nanopore technology. *Nat Commun* **10:** 1869. doi:10.1038/s41467-019-09637-5

Burge S, Parkinson GN, Hazel P, Todd AK, Neidle S. 2006. Quadruplex DNA: sequence, topology and structure. *Nucleic Acids Res* **34:** 5402–5415. doi:10.1093/nar/gkl655

Cer RZ, Donohue DE, Mudunuri US, Temiz NA, Loss MA, Starner NJ, Halusa GN, Volfovsky N, Yi M, Luke BT, et al. 2013. Non-B DB v2.0: a database of predicted non-B DNA-forming motifs and its associated tools. *Nucleic Acids Res* **41:** D94–D100. doi:10.1093/nar/gks955

Cook RD, Sanford W. 1982. *Residuals and influence in regression*. Chapman and Hall, New York.

Daniel B, Deamer DW. 2019. *Nanopore sequencing: an introduction*. World Scientific, Singapore.

Dohm JC, Lottaz C, Borodina T, Himmelbauer H. 2008. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res* **36:** e105. doi:10.1093/nar/gkn425

Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, et al. 2009. Real-time DNA sequencing from single polymerase molecules. *Science* **323:** 133–138. doi:10.1126/science.1162986

Fungtammasan A, Ananda G, Hile SE, Su MS-W, Sun C, Harris R, Medvedev P, Eckert K, Makova KD. 2015. Accurate typing of short tandem repeats from genome-wide sequencing data and its applications. *Genome Res* **25:** 736–749. doi:10.1101/gr.185892.114

Garrison E, Marth G. 2012. Haplotype-based variant detection from short-read sequencing. arXiv:1207.3907 [q-bio.GN].

Ghosh A, Bansal M. 2003. A glossary of DNA structures from A to Z. *Acta Crystallogr D Biol Crystallogr* **59:** 620–626. doi:10.1107/s0907444903003251

Guiblet WM, Cremona MA, Cechova M, Harris RS, Kejnovská I, Kejnovsky E, Eckert K, Chiaromonte F, Makova KD. 2018. Long-read sequencing technology indicates genome-wide effects of non-B DNA on polymerization speed and error rate. *Genome Res* **28:** 1767–1778. doi:10.1101/gr.241257.118

Hänsel-Hertsch R, Beraldi D, Lensing SV, Marsico G, Zyner K, Parry A, Di Antonio M, Pike J, Kimura H, Narita M, et al. 2016. G-quadruplex structures mark human regulatory chromatin. *Nat Genet* **48:** 1267–1272. doi:10.1038/ng.3662

Hile SE, Eckert KA. 2004. Positive correlation between DNA polymerase α-primase pausing and mutagenesis within polypyrimidine/polypurine microsatellite sequences. *J Mol Biol* **335:** 745–759. doi:10.1016/j.jmb.2003.10.075

Hile SE, Wang X, Lee MYWT, Eckert KA. 2012. Beyond translesion synthesis: polymerase κ fidelity as a potential determinant of microsatellite stability. *Nucleic Acids Res* **40:** 1636–1647. doi:10.1093/nar/gkr889

Htun H, Dahlberg J. 1988. Single strands, triple strands, and kinks in H-DNA. *Science* **241:** 1791–1796. doi:10.1126/science.3175620

Jain A, Wang G, Vasquez KM. 2008. DNA triple helices: biological consequences and therapeutic potential. *Biochimie* **90:** 1117–1130. doi:10.1016/j.biochi.2008.02.011

Jain A, Bacolla A, Chakraborty P, Grosse F, Vasquez KM. 2010. Human DHX9 helicase unwinds triple-helical DNA structures. *Biochemistry* **49:** 6992–6999. doi:10.1021/bi100795m

Jain M, Olsen HE, Paten B, Akeson M. 2016. Erratum to: The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol* **17:** 256. doi:10.1186/s13059-016-1122-x

Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, Tyson JR, Beggs AD, Dilthey AT, Fiddes IT, et al. 2018. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat Biotechnol* **36:** 338–345. doi:10.1038/nbt.4060

Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP, et al. 2020. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581:** 434–443. doi:10.1038/s41586-020-2308-7

Karst SM, Ziels RM, Kirkegaard RH, Sørensen EA, McDonald D, Zhu Q, Knight R, Albertsen M. 2021. High-accuracy long-read amplicon sequences using unique molecular identifiers with Nanopore or PacBio sequencing. *Nat Methods* **18:** 165–169. doi:10.1038/s41592-020-01041-y

Kelkar YD, Eckert KA, Chiaromonte F, Makova KD. 2011. A matter of life or death: how microsatellites emerge in and vanish from the human genome. *Genome Res* **21:** 2038–2048. doi:10.1101/gr.122937.111

Kishikawa T, Momozawa Y, Ozeki T, Mushiroda T, Inohara H, Kamatani Y, Kubo M, Okada Y. 2019. Empirical evaluation of variant calling accuracy using ultra-deep whole-genome sequencing data. *Sci Rep* **9:** 1784. doi:10.1038/s41598-018-38346-0

Koo H-S, Wu H-M, Crothers DM. 1986. DNA bending at adenine thymine tracts. *Nature* **320:** 501–506. doi:10.1038/320501a0

Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, Morgan MT, Carey VJ. 2013. Software for computing and annotating genomic ranges. *PLoS Comput Biol* **9:** e1003118. doi:10.1371/journal.pcbi.1003118

Lee H, Schatz MC. 2012. Genomic dark matter: the reliability of short read mapping illustrated by the genome mappability score. *Bioinformatics* **28:** 2097–2105. doi:10.1093/bioinformatics/bts330

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34:** 3094–3100. doi:10.1093/bioinformatics/bty191

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25:** 2078–2079. doi:10.1093/bioinformatics/btp352

Logsdon GA, Vollger MR, Eichler EE. 2020. Long-read human genome sequencing and its applications. *Nat Rev Genet* **21:** 597–614. doi:10.1038/s41576-020-0236-x

Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, Zhao M, Chennagiri N, Nordenfelt S, Tandon A, et al. 2016. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538:** 201–206. doi:10.1038/nature18964

Metzker ML. 2010. Sequencing technologies: the next generation. *Nat Rev Genet* **11:** 31–46. doi:10.1038/nrg2626

Mirkin EV, Mirkin SM. 2007. Replication fork stalling at natural impediments. *Microbiol Mol Biol Rev* **71:** 13–35. doi:10.1128/MMBR.00030-06

Nag DK, Petes TD. 1991. Seven-base-pair inverted repeats in DNA form stable hairpins in vivo in *Saccharomyces cerevisiae*. *Genetics* **129:** 669–673. doi:10.1093/genetics/129.3.669

Nielsen R, Paul JS, Albrechtsen A, Song YS. 2011. Genotype and SNP calling from next-generation sequencing data. *Nat Rev Genet* **12:** 443–451. doi:10.1038/nrg2986

Quail MA, Otto TD, Gu Y, Harris SR, Skelly TF, McQuillan JA, Swerdlow HP, Oyola SO. 2012. Optimal enzymes for amplifying sequencing libraries. *Nat Methods* **9:** 10–11. doi:10.1038/nmeth.1814

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26:** 841–842. doi:10.1093/bioinformatics/btq033

R Core Team. 2022. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/.

Sahakyan AB, Chambers VS, Marsico G, Santner T, Di Antonio M, Balasubramanian S. 2017. Machine learning model for sequence-driven DNA G-quadruplex formation. *Sci Rep* **7:** 14535. doi:10.1038/s41598-017-14017-4

Schirmer M, D'Amore R, Ijaz UZ, Hall N, Quince C. 2016. Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. *BMC Bioinformatics* **17:** 125. doi:10.1186/s12859-016-0976-y

Sen D, Gilbert W. 1988. Formation of parallel four-stranded complexes by guanine-rich motifs in DNA and its implications for meiosis. *Nature* **334:** 364–366. doi:10.1038/334364a0

Shafer ABA, Peart CR, Tusso S, Maayan I, Brelsford A, Wheat CW, Wolf JBW. 2017. Bioinformatic processing of RAD-seq data dramatically impacts downstream population genetic inference. *Methods in Ecol Evol* **8:** 907–917. doi:10.1111/2041-210x.12700

Shin S-I, Ham S, Park J, Seo SH, Lim CH, Jeon H, Huh J, Roh T-Y. 2016. Z-DNA-forming sites identified by ChIP-Seq are associated with actively transcribed regions in the human genome. *DNA Res* **23:** 477–486. doi:10.1093/dnares/dsw031

Sinden RR, Pytlos-Sinden MJ, Potaman VN. 2007. Slipped strand DNA structures. *Front Biosci* **12:** 4788–4799. doi:10.2741/2427

Singleton CK, Klysik J, Stirdivant SM, Wells RD. 1982. Left-handed Z-DNA is induced by supercoiling in physiological ionic conditions. *Nature* **299:** 312–316. doi:10.1038/299312a0

Slatkin M, Racimo F. 2016. Ancient DNA and human history. *Proc Natl Acad Sci USA* **113:** 6380–6387. doi:10.1073/pnas.1524306113

Stein M, Hile SE, Weissensteiner MH, Lee M, Zhang S, Kejnovský E, Kejnovská I, Makova KD, Eckert KA. 2022. Variation in G-quadruplex sequence and topology differentially impacts human DNA polymerase fidelity. *DNA Repair (Amst)* **119:** 103402. doi:10.1016/j.dnarep.2022.103402

Stoler N, Nekrutenko A. 2021. Sequencing error profiles of Illumina sequencing instruments. *NAR Genom Bioinform* **3:** lqab019. doi:10.1093/nargab/lqab019

Tabangin ME, Woo JG, Martin LJ. 2009. The effect of minor allele frequency on the likelihood of obtaining false positives. *BMC Proc* **3 Suppl 7:** S41. doi:10.1186/1753-6561-3-S7-S41

Wagner J, Olson ND, Harris L, Khan Z, Farek J, Mahmoud M, Stankovic A, Kovacevic V, Yoo B, Miller N, et al. 2022. Benchmarking challenging small variants with linked and long reads. *Cell Genom* **2:** 100128. doi:10.1016/j.xgen.2022.100128

Wainschtein P, Jain D, Zheng Z, TOPMed Anthropometry Working Group, NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium, Adrienne Cupples L, Shadyab AH, McKnight B, Shoemaker BM, et al. 2022. Assessing the contribution of rare variants to complex trait heritability from whole-genome sequence data. *Nat Genet* **54:** 263–273. doi:10.1038/s41588-021-00997-7

Wang G, Vasquez KM. 2014. Impact of alternative DNA structures on DNA damage, DNA repair, and genetic instability. *DNA Repair (Amst)* **19:** 143–151. doi:10.1016/j.dnarep.2014.03.017

Wang AH-J, Quigley GJ, Kolpak FJ, Crawford JL, van Boom JH, van der Marel G, Rich A. 1979. Molecular structure of a left-handed double helical DNA fragment at atomic resolution. *Nature* **282:** 680–686. doi:10.1038/282680a0

Wickham H. 2011. ggplot2. *Wiley Interdiscip Rev Comput Stat* **3:** 180–185. doi:10.1002/wics.147

Zhao J, Bacolla A, Wang G, Vasquez KM. 2010. Non-B DNA structure-induced genetic instability and evolution. *Cell Mol Life Sci* **67:** 43–62. doi:10.1007/s00018-009-0131-2

Zook JM, Catoe D, McDaniel J, Vang L, Spies N, Sidow A, Weng Z, Liu Y, Mason CE, Alexander N, et al. 2016. Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci Data* **3:** 160025. doi:10.1038/sdata.2016.25

Received November 9, 2022; accepted in revised form May 4, 2023.