Supplementary Figures

Figure S1. Heatmap of fold changes in aggregate error rates. Fold changes of different aggregate error rates (total number of mismatches divided by total number of aligned nucleotides) for single-nucleotide (SNM), insertion (INS), and deletion (DEL) errors. The left and right panels correspond to the moderately and stringently filtered motif sets, respectively, whereas rows correspond to Illumina, HiFi, and ONT technology. Shades of red and green indicate higher and lower error rate in non-B motifs than in controls, respectively. Note that these values are based on the aggregate, and not per-motif (and per-control) error rates (the latter are shown in the main manuscript).

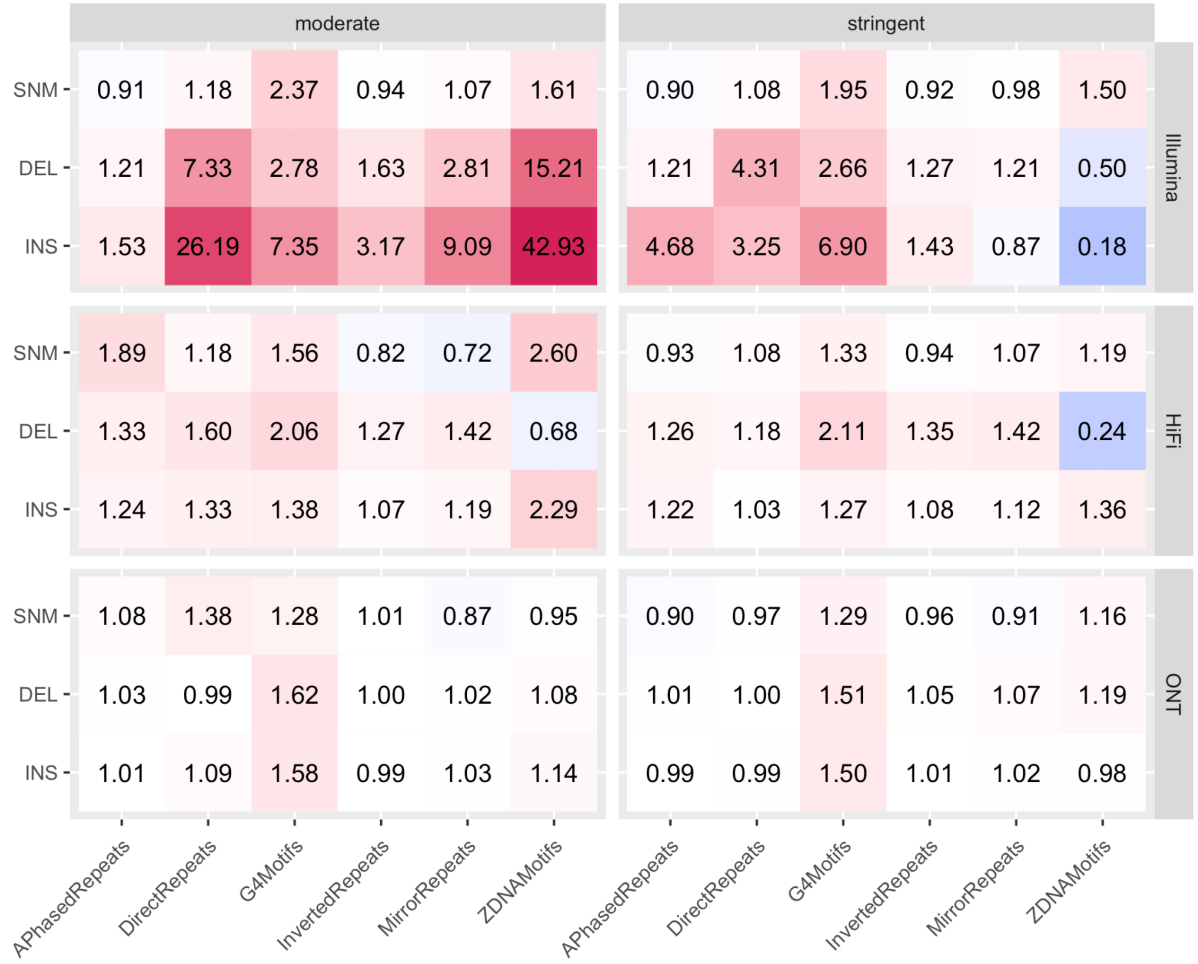


Figure S2. Single-nucleotide mismatch error rates in overlapping regions of paired-end reads. Shown are boxplots of SNM error rates derived from mismatches between pairs in overlapping Illumina read pairs. 95% confidence intervals were computed using the normal approximation of the binomial. To calculate the SNM error rate, the total number of mismatches was divided by the total number of nucleotides in the overlap. To minimize the effect of position in the read on the error rate, only the middle 10% of the reads were used. See Methods and Supplementary Table S5 for details.

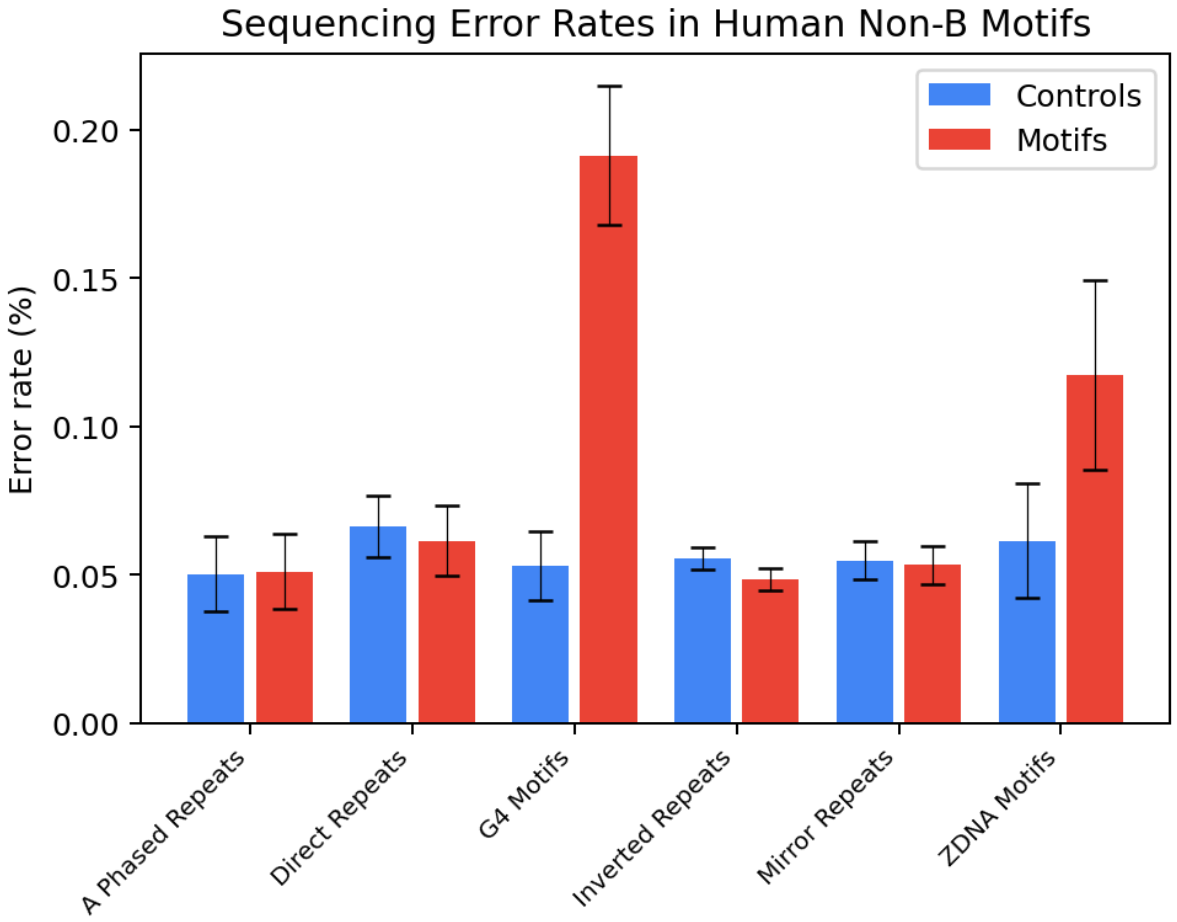
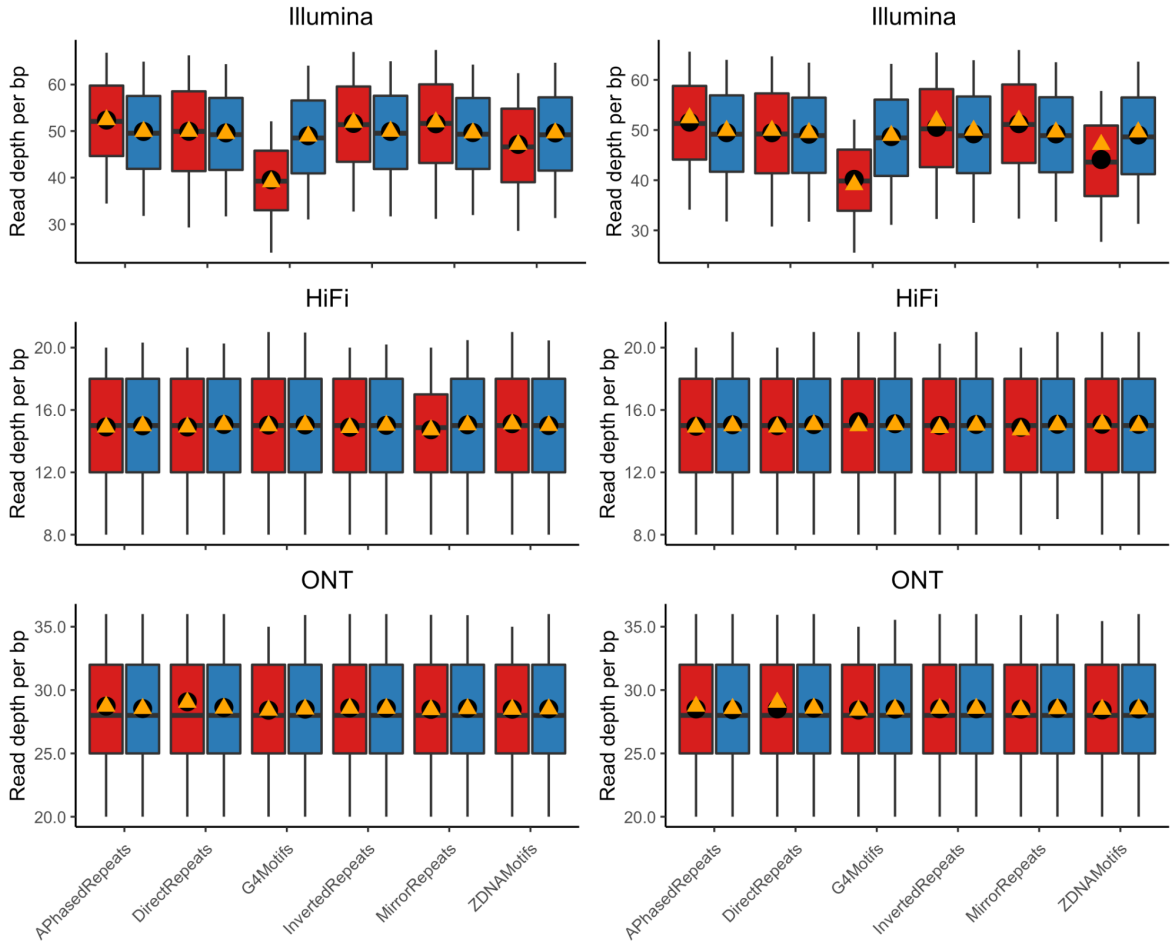


Figure S3. Read depth in non-B motifs. (A) Boxplots of per-bp read depth. Values above the 90th percentile were excluded from the figure to enable better visualization. The left panel shows results for the moderately filtered motif set, the right panel for the stringently filtered set, and the three rows correspond to the different technologies (Illumina, HiFi, and ONT). Black dots show values for per-motif means, orange triangles show overall error rates (sum of all aligned nucleotides divided by the total length of motifs / controls). **(B)** Heat map plot with fold changes of per-motif means of read depth. Red shades indicate lower values in non-B motifs than in controls, green shades indicate higher read depths in non-B motifs than in controls, with values also printed in rectangles. Values in bold represent fold-changes for which per-motif means were significantly different between motif and control (t-test p-values corrected for multiple testing). Left and right panels correspond to moderately and stringently filtered motif sets, respectively, and rows correspond to Illumina, HiFi, and ONT technologies.

A



B

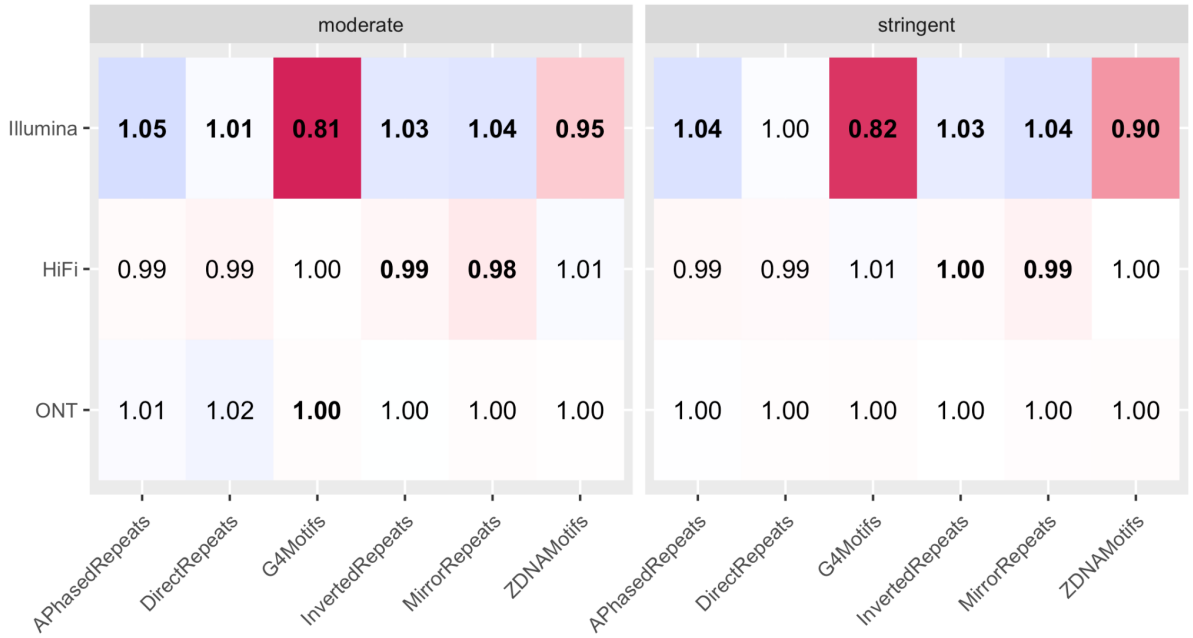
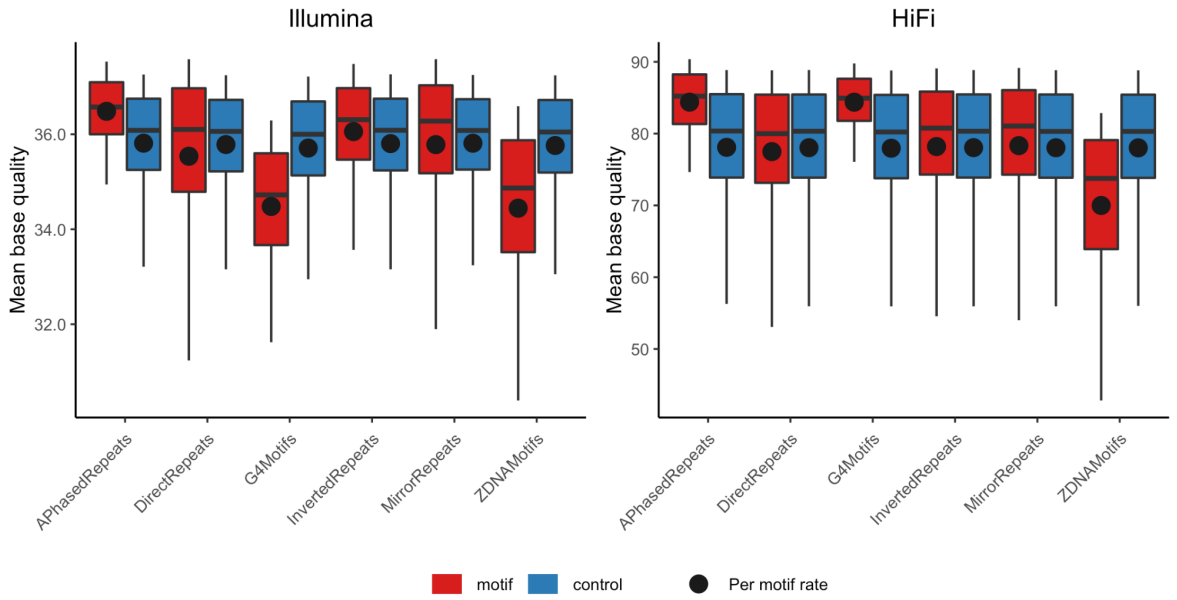

moderate
stringent
Illumina
HiFi
ONT
1.05 1.01 0.81 1.03 1.04 0.95
0.99 0.99 1.00 0.99 0.98 1.01
1.01 1.02 1.00 1.00 1.00 1.00
1.04 1.00 0.82 1.03 1.04 0.90
0.99 0.99 1.01 1.00 0.99 1.00
1.00 1.00 1.00 1.00 1.00 1.00
APhasedRepeats
DirectRepeats
G4Motifs
InvertedRepeats
MirrorRepeats
ZDNAMotifs

Figure S4. Base quality in non-B motifs. (A) Boxplots of average base quality. Values above the 90th percentile are excluded from the figure. Panels show results for Illumina and HiFi, respectively, and black dots show values for per-motif means. **(B)** Heat map plot with fold changes of per-motif means of base quality. Red shades indicate lower values in non-B motifs than in controls, green shades indicate higher base quality in non-B motifs than in controls, with values also printed in rectangles. Values in bold represent fold-changes for which per-motif means were significantly different between motif and control (t-test p-values corrected for multiple testing). Since we filtered for base quality in the stringently filtered motif set, we only show the results for the moderately filtered set.

A



B

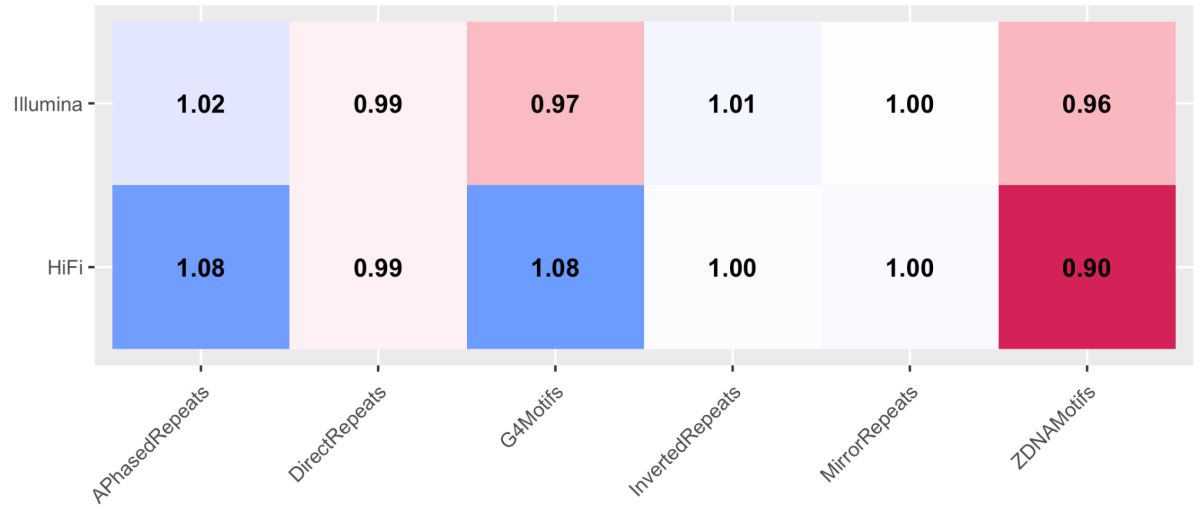


Figure S5. False-positive doubleton SNVs in non-B motifs. (A) Scenarios with different technologies corresponding to rows and numbers of diploid individuals (1000 and 10,000) corresponding to columns. The average read depth per haploid genome is plotted on the x-axis**,** whereas the expected number of false-positive SNVs due to errors that occur at least on two chromosomes among the sampled individuals (doubletons) is shown on the y-axis. The solid line corresponds to the cumulative number of all false-positive SNVs across non-B types, and the dashed line to the number of false-positive SNVs in an equally long stretch of B-DNA. **(B)** Expected false-positive doubleton 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 only doubletons considered (A).

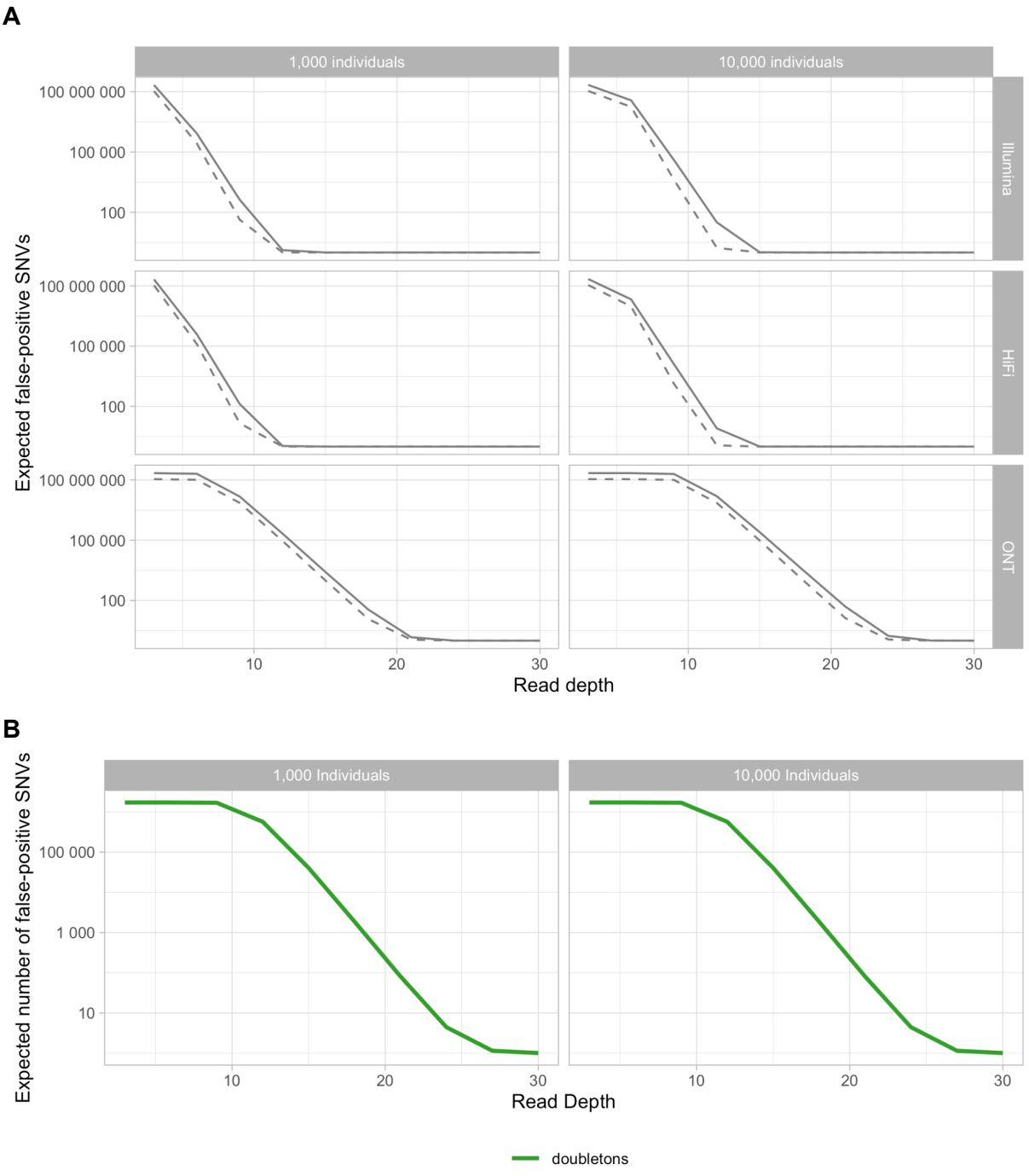
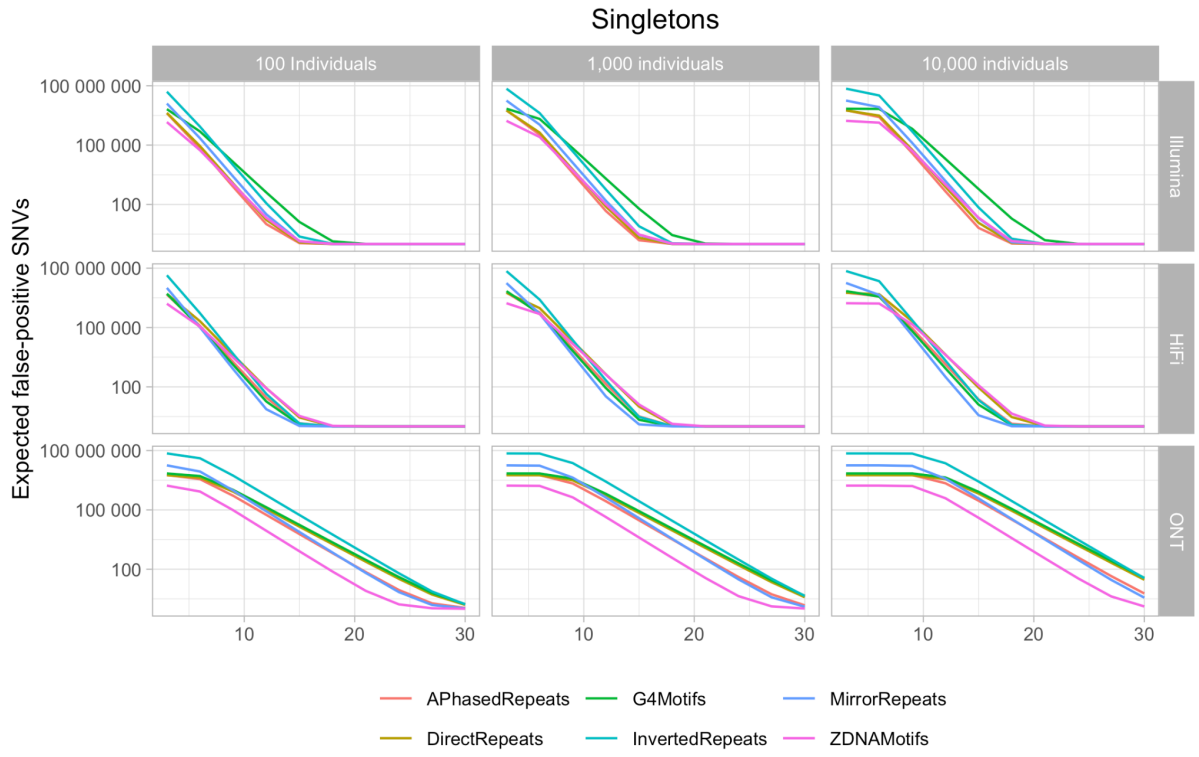
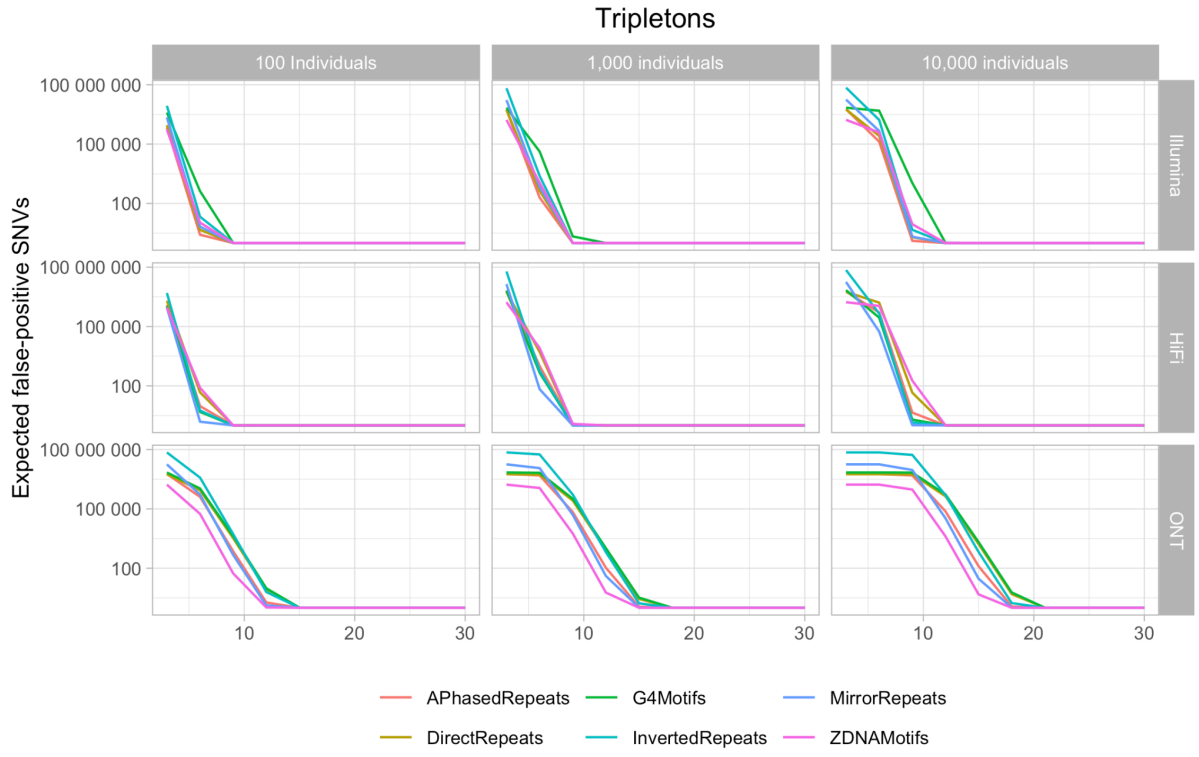


Figure S6. False-positive SNVs in non-B motifs. Shown are numbers of expected false-positive SNVs (y-axis) based on the probabilistic model and the SNM error rates derived from non-B motifs, for different haploid read depths (x-axis). For this analysis, we considered all bases annotated as a non-B motif of a particular type in the genome, i.e. the number of bases for each motif type differs among motif types. Different non-B motif types are shown in different colors. Columns correspond to different numbers of individuals, rows to the three technologies. (A), (B), and (C) show results for singleton, tripleton and 1% variant allele frequency cutoffs, respectively.

A



B


Tripletons
100 Individuals
1,000 individuals
10,000 individuals
Illumina
HiFi
ONT
Expected false-positive SNVs
100 000 000
100 000
100
10
20
30
APhasedRepeats
DirectRepeats
G4Motifs
InvertedRepeats
MirrorRepeats
ZDNAMotifs

C

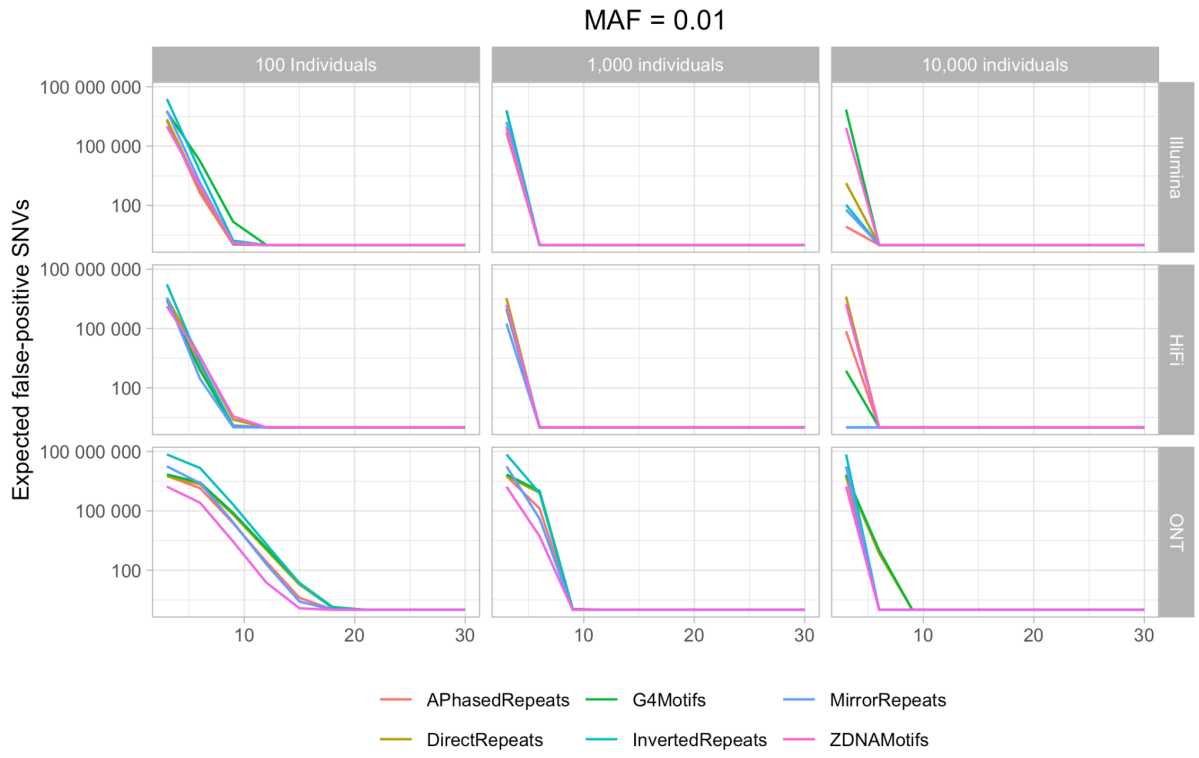

MAF = 0.01
100 Individuals
1,000 individuals
10,000 individuals
Illumina
HiFi
ONT
Expected false-positive SNVs
100 000 000
100 000
100
10
20
30
APhasedRepeats
DirectRepeats
G4Motifs
InvertedRepeats
MirrorRepeats
ZDNAMotifs

Figure S7. False-positive SNVs in non-B motifs, example of simulation results. Shown are numbers of expected false-positive SNVs in non-B motifs (red), and controls (blue), based on error rates in Illumina sequencing, and a minor variant frequency cutoff of one variant (singletons) in 100 diploid individuals. As opposed to results shown in Fig. 6, which only included the expected number of false positives, the boxplots of false positive values presented here are based on 100 Monte-Carlo simulations of each site.

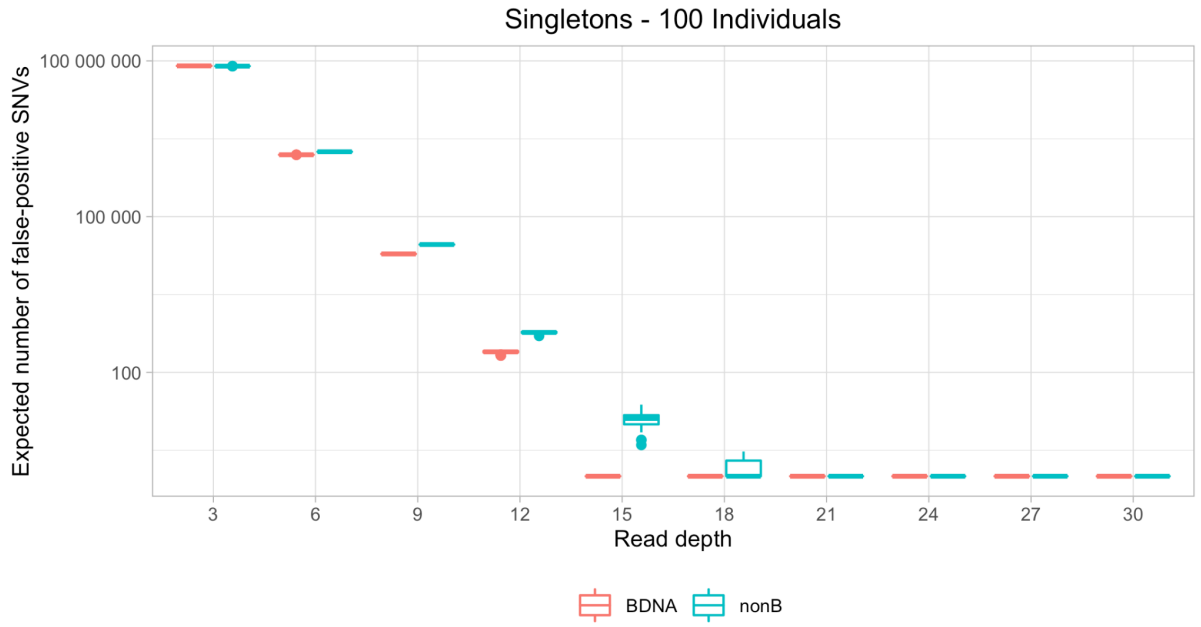


Figure S8. Empirical estimation of false-positive SNVs in non-B motifs and controls. Shown are proportions of false-positive SNVs calculated by using the HG002 Illumina data set that was also employed in the error detection analysis. Read depth was subsampled to 3, 9, 15, and 30× (x-axis), and variants were identified using FreeBayes, with default parameters and a minimum mapping quality filter of 30. To estimate the proportion of false-positive SNVs in the combined non-B DNA annotation and the respective controls, we compared the called variants with two versions of the GIAB true variant set: v3.3.2 (which was also used in the main analysis, panel **A**) and v4.2.1 (the most recent version, panel **B**). Overall, proportions of false-positive SNVs start to decline from 15× on, with markedly higher proportions in the non-B motifs compared to the controls. Interestingly, such difference seems to be even more pronounced when using the newer true variant set, v4.2.1. Note that the increase in proportion of false-positives from 3× to 9× read depth is probably due to the fact that at read depth 3 several genomic regions are not covered by any read.

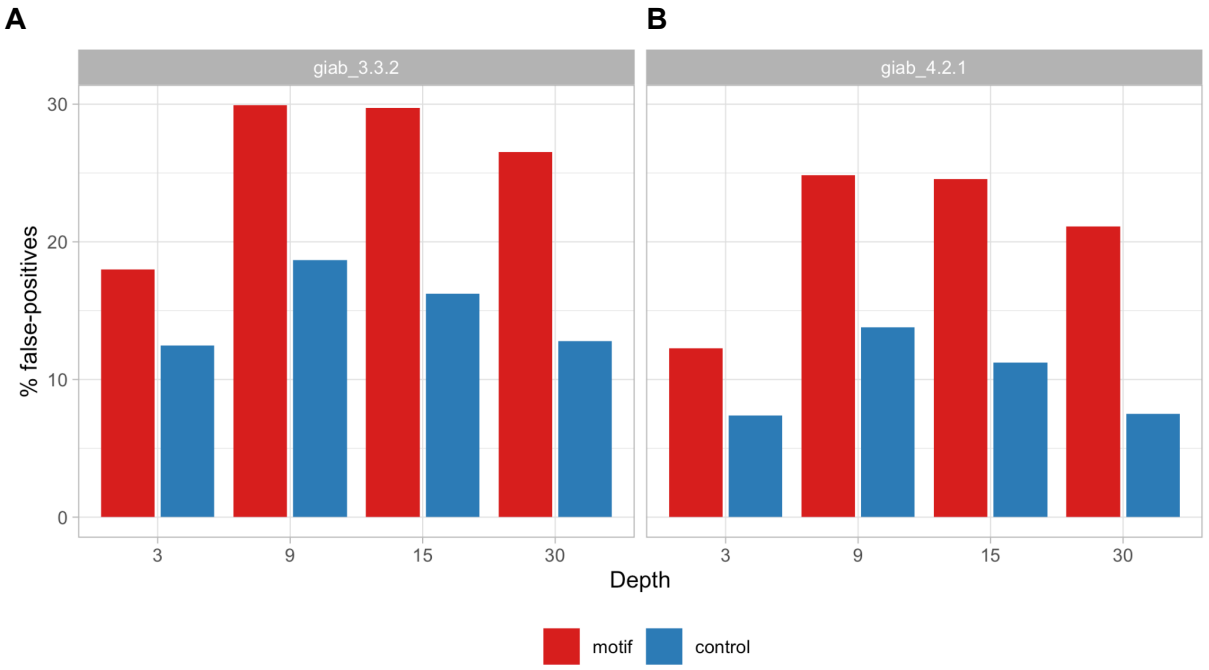


Figure S9. Genome-wide nucleotide substitution frequencies at G4 motifs and corresponding controls (for the stringent filtered data). The positions of nucleotide substitutions within motifs were scaled based on motif size. Stems are runs of guanines and loops are unspecified nucleotides between stems. For clarity of visualization, the Y-axis is displayed on a log scale. A comparison between all G4 loci (shown in blue) and control sequences (shown in black) for single-nucleotide polymorphism (SNP) frequencies based on the Simons Genome Diversity Project (Mallick et al. 2016) analyzed in (Guiblet et al. 2021).

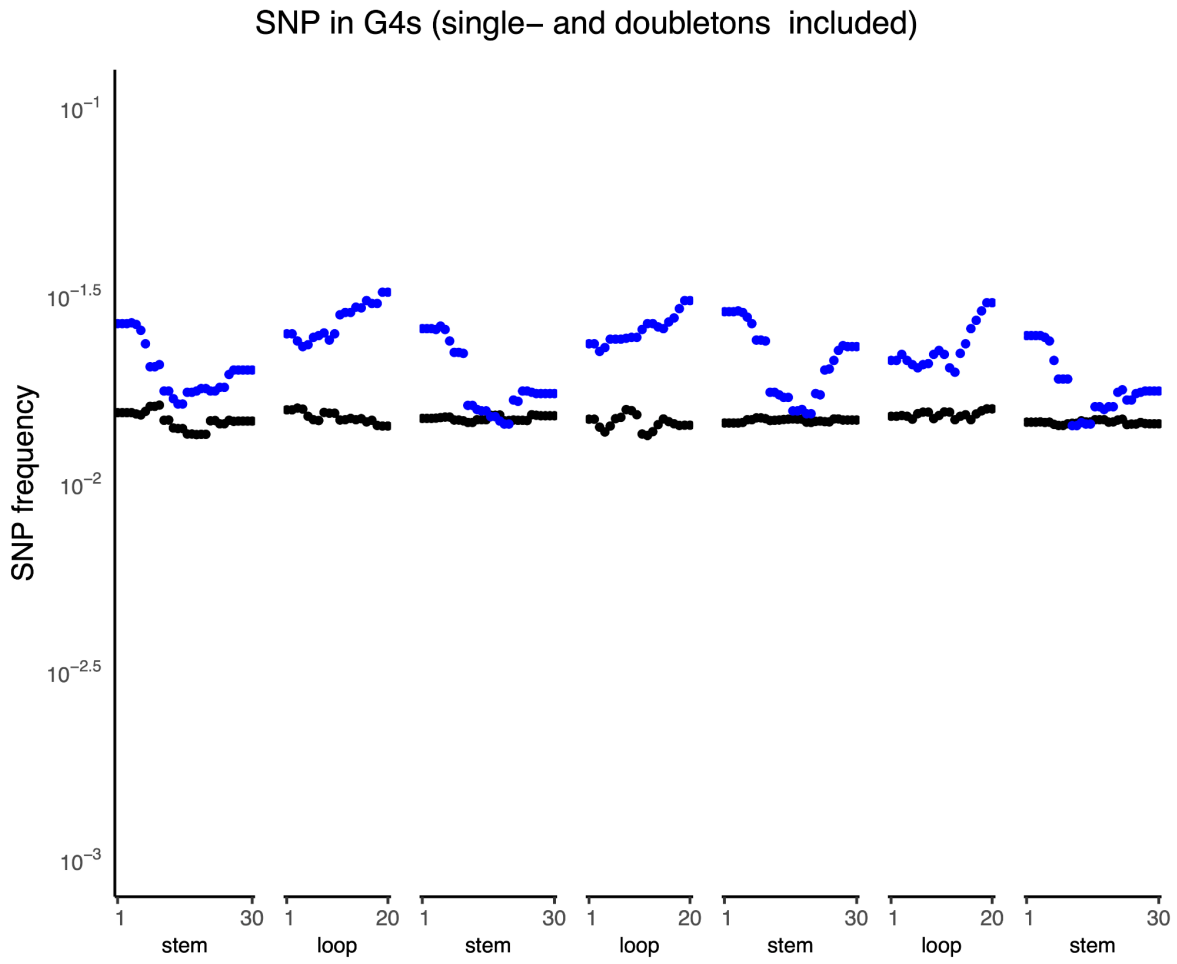
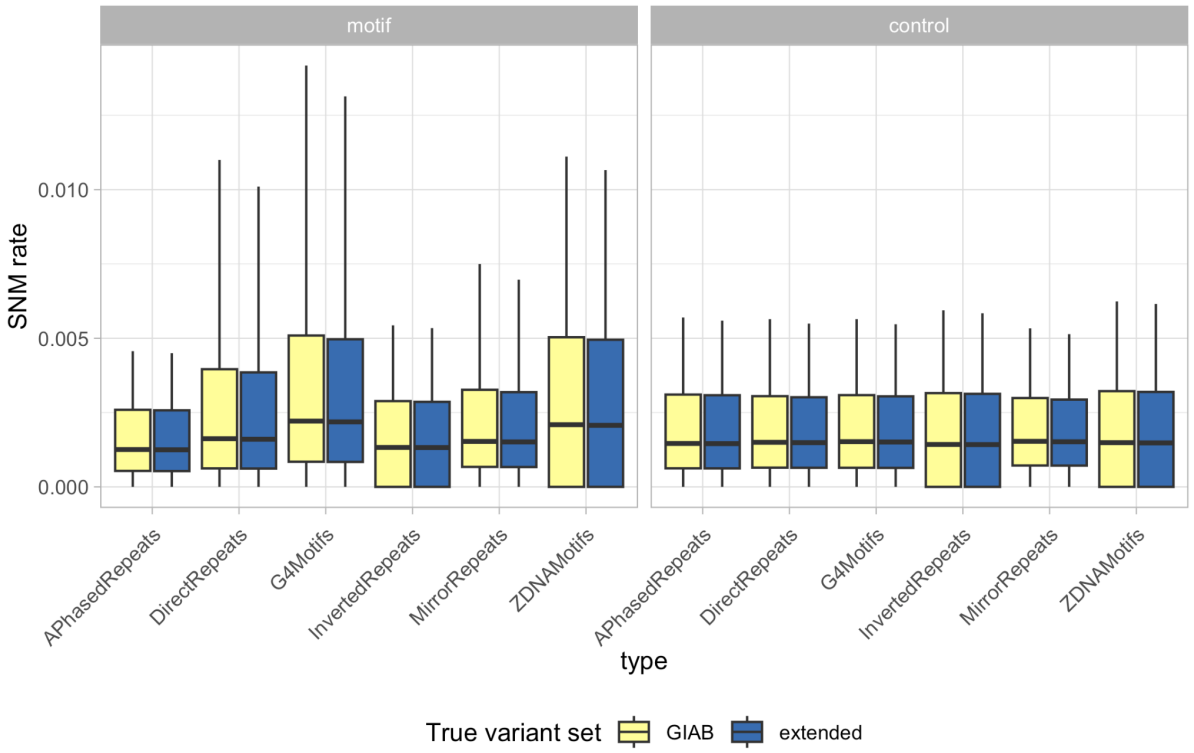


Figure S10. SNV error rates using an extended true variant set. To test whether using a more comprehensive true variant set extending the GIAB version 3.3.2 has an effect on the observed error rate patterns, we identified SNVs in the Illumina, HiFi, and ONT data for HG002 using *varscan* (Koboldt et al. 2009). The extended true variant set comprises all variants present in at least two of the four callsets (Illumina, HiFi, ONT, and GIAB v3.3.2). We then repeated the error detection analysis for chromosome 1 of the Illumina data. As shown in the figure, the SNV rates for both motifs and controls are remarkably similar between the GIAB v3.3.2 (yellow), and the extended (blue) true variant set. Therefore, we are confident that the use of solely the GIAB v3.3.2 true variant set does not influence our conclusions.



REFERENCES

Guiblet, Wilfried M., Marzia A. Cremona, Robert S. Harris, Di Chen, Kristin A. Eckert, Francesca Chiaromonte, Yi-Fei Huang, and Kateryna D. Makova. 2021. “Non-B DNA: A Major Contributor to Small- and Large-Scale Variation in Nucleotide Substitution Frequencies across the Genome.” *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkaa1269.

Koboldt, Daniel C., Ken Chen, Todd Wylie, David E. Larson, Michael D. McLellan, Elaine R. Mardis, George M. Weinstock, Richard K. Wilson, and Li Ding. 2009. “VarScan: Variant Detection in Massively Parallel Sequencing of Individual and Pooled Samples.” *Bioinformatics* 25 (17): 2283–85.

Mallick, Swapan, Heng Li, Mark Lipson, Iain Mathieson, Melissa Gymrek, Fernando Racimo, Mengyao Zhao, et al. 2016. “The Simons Genome Diversity Project: 300 Genomes from 142 Diverse Populations.” *Nature* 538 (7624): 201–6.

Schirmer, Melanie, Rosalinda D’Amore, Umer Z. Ijaz, Neil Hall, and Christopher Quince. 2016. “Illumina Error Profiles: Resolving Fine-Scale Variation in Metagenomic Sequencing Data.” *BMC Bioinformatics* 17 (March): 125.