Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Context-dependent DNA Polymerization Effects Can Masquerade as DNA Modification Signals

Yusuke Takahashi
The University of Tokyo: Tokyo Daigaku https://orcid.org/0000-0002-1640-1138
Massa Shoura
Stanford University School of Medicine https://orcid.org/0000-0003-3278-296X
Andrew Fire
Stanford University School of Medicine https://orcid.org/0000-0001-6217-8312
Shinichi Morishita (moris@edu.k.u-tokyo.ac.jp)
University of Tokyo https://orcid.org/0000-0002-6201-8885

Research article

Keywords: DNA polymerization, DNA modification, non-B DNA, whole genome amplification, single-molecule real-time (SMRT) sequencing, DNA N6-methyladenine

Posted Date: May 24th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-549532/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Abstract

Background

Single molecule measurements of DNA polymerization kinetics provide a sensitive means to detect both secondary structures in DNA and deviations from primary chemical structure as a result of modified bases. In one approach to such analysis, deviations can be inferred by monitoring the behavior of DNA polymerase using single-molecule, real-time sequencing with zero-mode waveguide. This approach measures the time between fluorescence pulse signals from consecutive nucleosides incorporated during DNA replication, called the interpulse duration (IPD).

Results

In this paper we present an analysis of loci with high IPDs in two genomes, a bacterial genome (E. coli) and a eukaryotic genome (C. elegans). To distinguish the potential effects of DNA modification on DNA polymerization speed, we paired an analysis of native genomic DNA with whole-genome amplified (WGA) material in which DNA modifications were effectively removed. Modification sites for E. coli are known and we observed the expected IPD shifts at these sites in the native but not WGA samples. For C. elegans, such differences were not observed. Instead, we found a number of novel sequence contexts where IPDs were raised relative to the average IPDs for each of the four nucleotides, but for which the raised IPD was present in both native and WGA samples.

Conclusion

The latter results argue strongly against DNA modification as the underlying driver for high IPD segments for C. elegans, and provide a framework for separating effects of DNA modification from context-dependent DNA polymerase kinetic patterns inherent in underlying DNA sequence for a complex eukaryotic genome.

Full Text

Due to technical limitations, full-text HTML conversion of this manuscript could not be completed. However, the manuscript can be downloaded and accessed as a PDF.

Tables

Table 1: Mean read length and average read coverage per strand in each sample.

	Replicate 1 /WGA	Replicate 2 /WGA	Replicate 1 /native	Replicate 2 /native
Mean read length (bp)	146	268	1,096	1,184
Average read coverage in the *C. elegans* genome	12.9	21.6	41.8	45.1
Average read coverage in the *E. Coli* genome	1.05	2.19	1.70	1.22

Figures

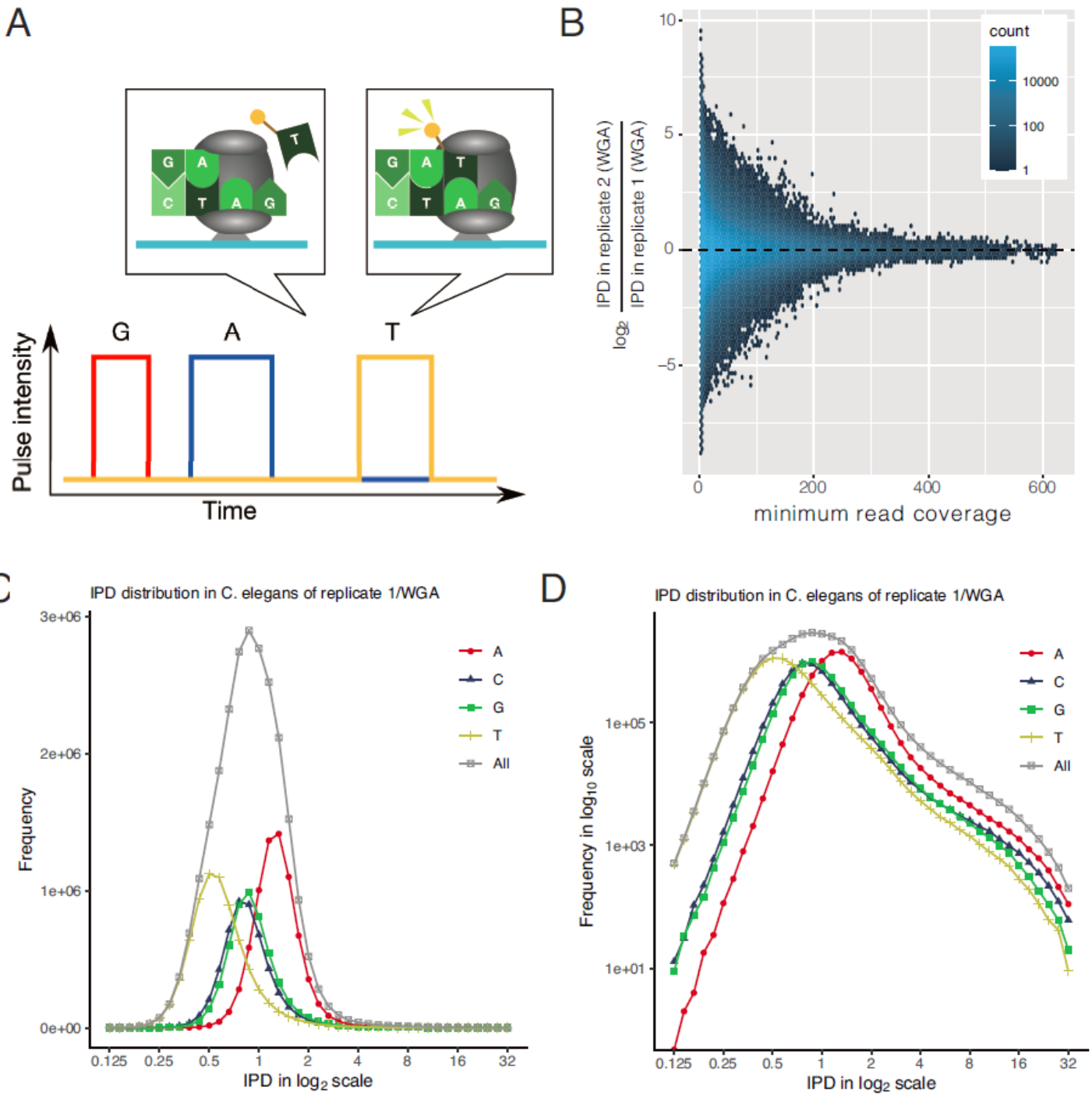

Figure 1

(A) The figure illustrates how a zero-mode waveguide monitor fluorescence signals from labeled nucleotides incorporated during DNA replication of the single-strand template of 5′-GATC- 3′. The interpulse duration (IPD), the time between pulse signals from consecutive nucleosides, is useful in detecting methylated or damaged nucleotides in bacterial genomes because of slower incorporation of nucleotides by DNA polymerase. (B) Hexbin plot of the logarithmic scale ratio of IPD in replicate 2/WGA to IPD in replicate 1/WGA at each base for the minimum read coverage shown in the x-axis in each strand

of the WGA samples. The IPDs fluctuate remarkably between the two WGA biological replicates when the minimum read coverage is low. (C) The frequency distribution of the IPDs in the x-axis (in log2 scale) of all bases and of individual four bases. (D) The y-axis of Figure C is represented in log10 scale to highlight the frequencies of high IPDs > 2.

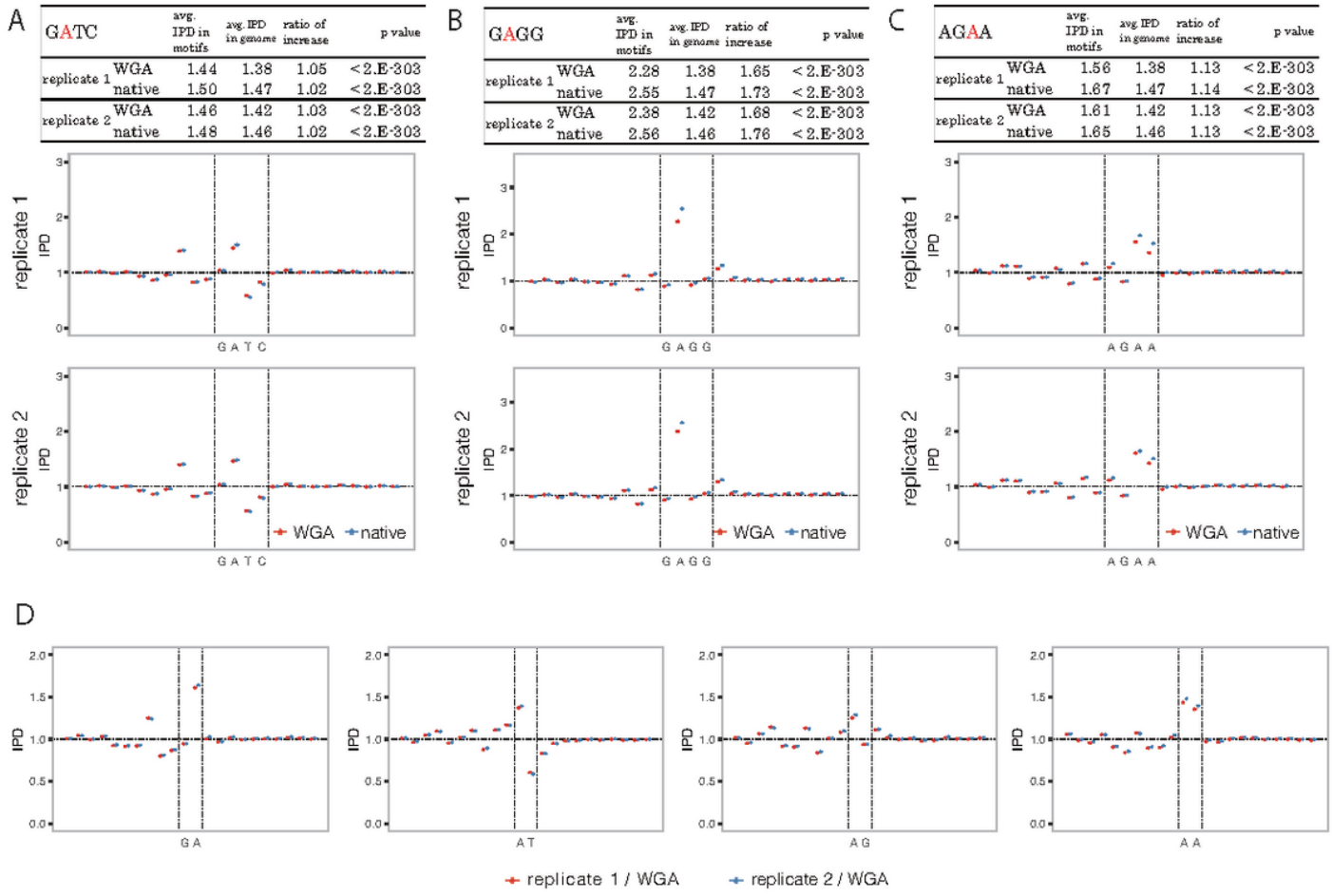

Figure 2

(A-C) Concordance between the IPDs in the WGA (colored red) and native (blue) samples (replicate 1 in the 2nd row, replicate 2 in the 3rd row). The top table shows the statistics of the focal nucleotide with the maximum IPD (colored red) in each motif; namely, the average IPD of the focal nucleotide in all motif occurrences, the average IPD in the entire C. elegans genome, and the ratio of increase, the ratio of the average IPD in motif occurrences to that in the genome. The significance of the ratio of increase is confirmed by comparing the frequency distributions of the IPDs using Wilcoxon's ranksum test (p-values shown in the last columns). The middle and bottom charts show the IPD distributions represented by an error bar plot in the motifs and their surrounding 10 nucleotides in the x-axis. Nearly identical IPD distributions are obtained from the two biological replicates (0.99 < Pearson's correlation coefficient; Supplementary Figure 7). Figure A displays GATC, where N6- methyladenine is prevalent in E. Coli. Figures B and C show motifs that were reported to have N6- methyladenine in C. elegans. (D) IPD distributions in the two WGA replicates around the four 2-mer motifs (GA, AT, AG, and AA) that have adenines and occur in the 4-mer motifs, GATC, GAGG, and AGAA in Figures A-C. The A's average IPD in GA are higher than

those of the other three 2-mer motifs, and almost concord with the average IPDs of A in GATC and AGAA, but is much smaller than the A's average IPD in GAGG.

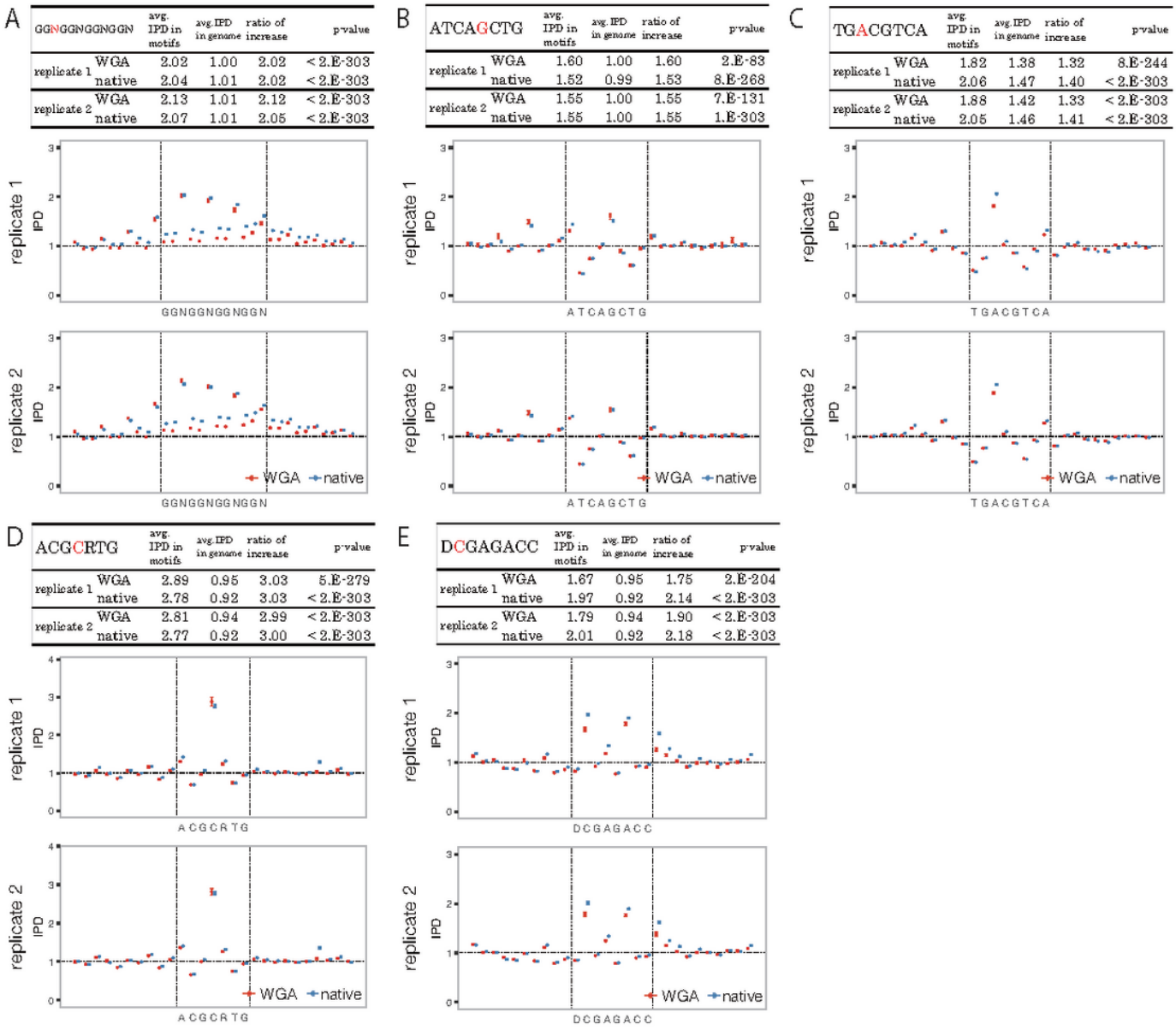

Figure 3

Similarly to Figure 2, Figures A-E show motifs that have one or more bases with extremely high IPD. Figures A, B, and C show motif examples in non-B DNA. (GGN)4 in A may form Gquadruplexes that are associated with polymerization slowdown. AT(CAG)(CTG) in B and (TGAC)(GTCA) in C have quasi-palindromes that are pairs of reverse-complementary sequences in parentheses and might induce cruciform DNA structures associated with polymerization acceleration. Figures D and E show motifs with extreme IPDs at cytosines.

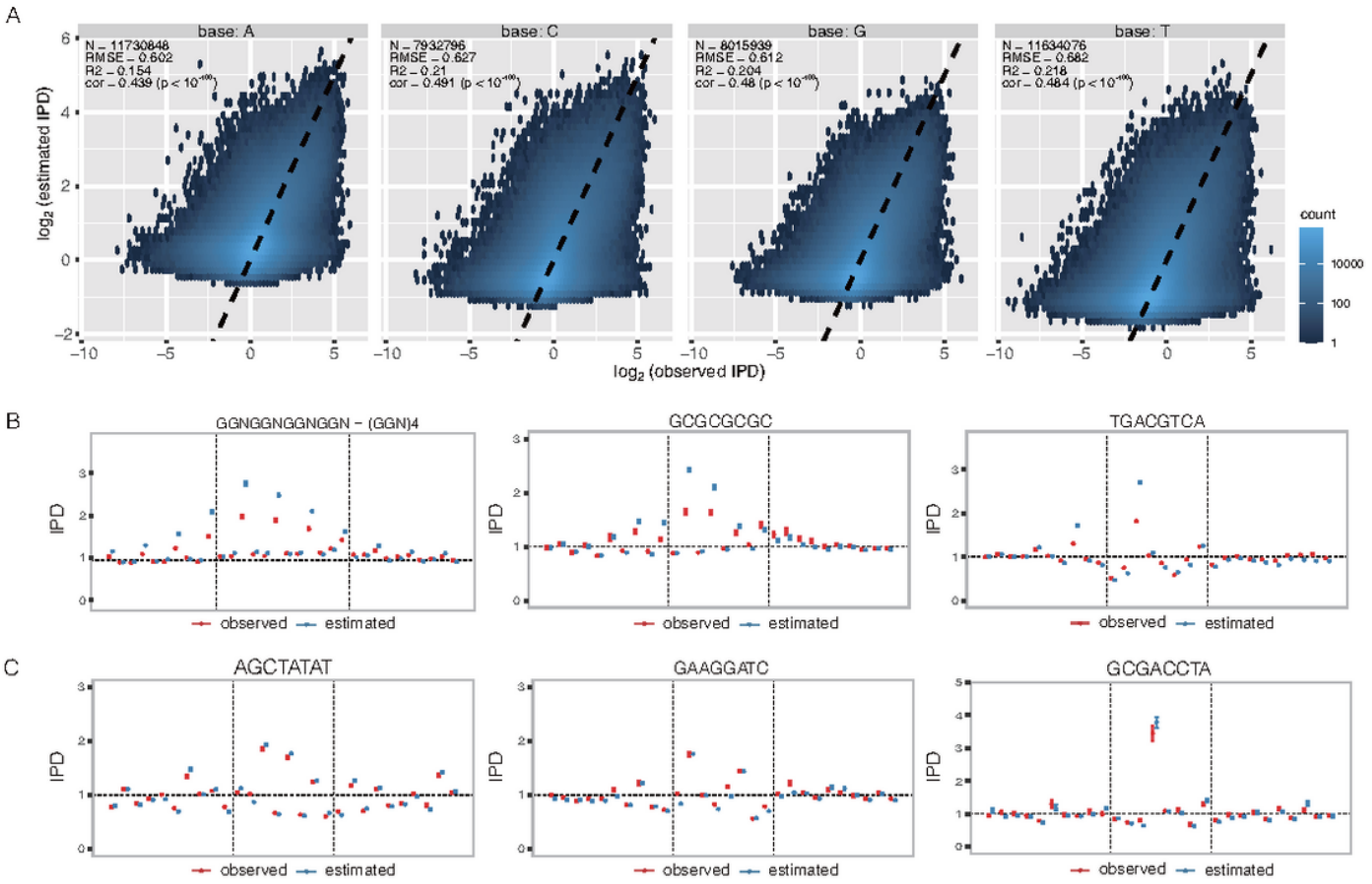

Figure 4

Discrepancy between the observed IPDs in the replicate 1/WGA sample and predicted IPDs in C. elegans. (A) Hexbin plot comparing the IPDs (x-axis) observed in the replicate 1/WGA sample with those estimated (y-axis)) using the PacBio software for each type of base. Values are shown using a logarithmic scale. Inside each plot, N, RMSE, R2, and cor represent the number of bases, root-meansquare error, R2 (coefficient of determination), and Pearson's correlation coefficient with the associated p-value. (B) Large discrepancies between the observed and estimated IPDs are seen at the bases with extreme IPDs in the three motifs; namely, at Ns in (GGN)4, Cs in GCGCGCGC, and A in TGACGTCA. (C) Meanwhile, predictions are almost consistent with observations in several motifs. Supplementary Figure 12 and 13 show the observed and predicted IPDs around all motifs in the four samples.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementaryFiguresTables.v13.pdf