

## Research Article

# IMPACT OF FLUORESCENT DYES ON MUTATIONS IN NEXT GENERATION SEQUENCING LIBRARY GENERATION

Vincent Butty, MD, PhD<sup>1,2,3</sup> , Pranav Patel<sup>4</sup>, Stefan Green, PhD<sup>5</sup> , Stuart Levine, PhD<sup>1,2,3</sup> 

<sup>1</sup> Biology, Massachusetts Institute of Technology, <sup>2</sup> Biological Engineering, Massachusetts Institute of Technology, <sup>3</sup> Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology, <sup>4</sup> NA, n6TEC, <sup>5</sup> Genomics and Microbiome Core Facility, Rush University Medical Center

Keywords: qPCR, NGS, Library Quantification, Mutation Signature, 16S

<https://doi.org/10.7171/001c.166072>

DNA labelling fluorescent dyes such as ethidium bromide have long been considered to be highly mutagenic during DNA replication. While recent studies have pushed back on this narrative, the intercalative nature of these dyes continues to raise the possibility that these dyes can induce mutations. The iconPCR instrument by n6tec uses fluorescent dyes to measure amplification in real time and to adjust cycling conditions. However, since this use of qPCR is preparative and not analytical, mutations introduced by fluorescent dyes would be propagated into the sequencing reaction. To address the impact of these dyes on downstream analyses, we have performed routine mutation calling as well as mutational signature analysis on samples amplified using the iconPCR in the presence of either SYBR or EvaGreen. Sequence analysis revealed very minimal impacts of dyes on the reactions, largely within the noise regimen with only subtle changes in mutation rates seen. Mutational signature analysis was unable to identify any key signatures assignable to the dyes in either substitutions or indel domains. The mutational impact of intercalating dyes during fluorescence-guided amplification is therefore minimal and can be disregarded in all but the most sensitive NGS applications.

## Introduction

Intercalating dyes have been used to detect and measure DNA for several decades.<sup>1</sup> These dyes fluoresce robustly in the context of double-stranded DNA, and the variable fluorescence led to the invention of quantitative PCR, which incorporates optical detection into the typical PCR reaction to measure the amount of fluorescence after each cycle. By measuring the cycle at which fluorescence crosses a set threshold, these instruments quantify the amount of input material included in each reaction. The intercalating mechanism of these dyes, inserting between adjacent base pairs and distorting the DNA helix, raises a direct concern regarding mutagenicity. Intercalating dyes have long been classified as suspected mutagens and possible carcinogens in laboratory safety documentation, a classification supported by early bacterial mutagenicity assays.<sup>2</sup> More recent studies have found that mutagenic potential requires metabolic activation and may be substantially lower than initially feared,<sup>3</sup> though the possibility that these dyes could introduce errors during DNA replication remains an open question.

The n6tec iconPCR (Individually Controlled PCR) instrument (n6tec, Pleasanton, CA, USA) takes quantitative PCR applications a step further. Rather than applying uniform cycling conditions across an entire plate, the instrument independently monitors and controls each of 96 wells in real time, halting individual reactions as they exit the exponential amplification phase or meet other user-specified conditions. This per-well control offers a significant advantage over conventional fixed-cycle thermocyclers in next-

generation sequencing (NGS) library preparation, where both over- and underamplification of libraries presents critical challenges. Underamplification yields insufficient material and may necessitate re-amplification, introducing both delay and variability. Over-amplification, particularly reactions driven into the PCR plateau phase, is associated with GC and composition bias, increased duplicate rates, chimera formation, and the generation of spurious higher-molecular-weight heteroduplex products, all of which can compromise library pooling and downstream sequencing analyses.<sup>4-6</sup>

The use of intercalating dyes in the iconPCR instrument (n6tec) therefore raises a specific concern: dye-induced errors introduced during preparative amplification would be propagated into the sequencing library and could confound downstream analyses. To assess this risk, we employed mutational signature analysis alongside conventional variant calling in a simple experimental system. Mutational signatures are characteristic combinations of mutation types arising from specific mutagenic processes; a mutagen acting through a defined mechanism, such as base intercalation, would be expected to produce a recognizable pattern of mutations attributable to that mechanism.<sup>7</sup> We therefore performed detailed analysis of reference samples prepared using the iconPCR instrument (n6tec) in the presence of intercalating dyes and compared to controls amplified without dyes. Bulk analysis revealed minimal changes in both single-nucleotide variation rates and insertion or deletion frequencies. Mutational signature analysis revealed minimal signatures not associated with known mechanisms. Thus, amplification with the iconPCR instrument (n6tec) should have minimal impact on next-generation sequenc-

ing (NGS) analyses except in the most sensitive applications.

## Materials and Methods

### Human Genomic DNA and Library Preparation

Human genomic DNA from Genome In A Bottle (GIAB) reference material NA12878 was obtained from the Coriell Institute (Camden, NJ, USA).<sup>8,9</sup> DNA was enzymatically fragmented to a target insert size of approximately 150–200 bp using Ultra II FS DNA Library Prep kit (New England Biolabs, Ipswich, MA, USA, catalog no. E7805). Following fragmentation, end repair, dA-tailing, and ligation of P5/P7 Illumina-compatible adapters (New England Biolabs) were performed per the manufacturer's protocol. Ligation products were purified using NEBNext Sample Purification Beads (New England Biolabs, catalog no. E6552A), at a 0.8X ratio.

### IconPCR Amplification

Amplification was performed on an iconPCR instrument (n6tec) using NEB Q5 High-Fidelity DNA Polymerase (New England Biolabs). 2.5 ng of ligated material was amplified in the presence of SYBR Green I (Thermo Fisher Scientific), EvaGreen (Biotium), or EvaGreen Plus (Biotium) at 1X (final concentration) alongside a no-dye control, for 10 cycles under the following conditions: initial denaturation at 98°C for 30 seconds, 10 cycles at 98°C for 10 seconds and 65°C for 75 seconds, final extension at 65°C for 5 minutes and hold at 4°C. Dye concentrations and cycle counts were intentionally set above the iconPCR instrument's (n6tec) standard stop point parameters to provide a stringent test of mutagenic potential. Amplified libraries were purified using NEBNext Sample Purification Beads (New England Biolabs), at a 0.8X ratio.

### Sequencing, Alignment, and Quality Control

Libraries were quantified using an Agilent Fragment Analyzer (Agilent Technologies, Santa Clara, CA, USA) and by qPCR and sequenced on an Element AVITI using 2 x 75 nt reads to a depth of approximately 100 million reads per sample (93–138M). Sequencing reads from the four conditions were aligned against GRCh38 using BWA-MEM (VN:0.7.17-r1188).<sup>10</sup> Quality control metrics were calculated (including number of aligned reads, multiply-mapping reads, and number of unique 20-mers in the top 10 million reads) and samples were checked for contamination against a collection of reference genomes.

### Variant Calling

Each sample was mapped against the GIAB GRCh38-based reference assembly GRCh38\_GIABv3\_no\_alt\_analyses\_set\_maskedGRC\_decoys\_MAP2K3\_KMT2C\_KCNJ18.fasta, obtained from <https://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/refer->

[ences/](https://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/refer-), using BWA-MEM (v. 0.7.12-r1039, -t 16 cores).<sup>10</sup> Resulting BAM files were sorted and indexed using SAMtools v.1.3. Variant calling was performed using BCFtools mpileup (v. 1.10.2+htslib-1.10.2) with parameters -B -x -d 1000000 -threads 8 -A -O v, against the above reference assembly.<sup>11,12</sup> VCF files were filtered against the GIAB benchmark VCF for sample NA12878\_HG001 (HG001\_GRCh38\_1\_22\_v4.2.1\_benchmark.vcf.gz, retrieved from [https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/NA12878\\_HG001/latest/GRCh38/](https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/NA12878_HG001/latest/GRCh38/)) using the BCFtools isec function (flags -c all -O v).<sup>9</sup> No additional filtering was performed. SNV rates were calculated as total observed variant bases divided by total bases sequenced; indel rates were calculated analogously.

### 16S rRNA Gene Amplicon Library Preparation

A 16S rRNA gene amplicon control was generated from the ZymoBIOMICS Microbial Community DNA Standard (Zymo Research, Irvine, CA, USA; catalog no. D6306) using a two-stage PCR protocol as described previously.<sup>13</sup> The V4–V5 region of the microbial small subunit rRNA gene was targeted using primers sIDTP5\_515F and sIDTP7\_926R (CTA-CACGACGCTCTTCCGATCTGTGYCAGCMGCCGCGGTAA and CAGACGTGTGCTCTTCCGATCTCCGYCAATTYMTT-TRAGTTT, respectively; underlined regions represent linker sequences), which carry 5' common sequence tags (sIDTP5 and sIDTP7) matching 3' sequences present in IDT xGen Amplicon UDI primers (Integrated DNA Technologies, Coralville, IA, USA). First-stage amplifications were performed in 10 µL reactions using repliQa HiFi ToughMix (Quantabio, Beverly, MA, USA), with template added at 5 ng/reaction, in the presence or absence of SYBR Green I Nucleic Acid Gel Stain (Thermo Fisher Scientific; catalog no. S7563) at 1X working concentration. Cycling conditions were 98°C for two minutes, followed by 16, 20, 24, or 28 cycles of 98°C for 10 seconds, 50°C for one second, and 68°C for one second. A second-stage amplification was then performed in 10 µL reactions using repliQa HiFi ToughMix (Quantabio), with each well receiving a unique dual-index primer pair (IDT xGen Amplicon UDI primer sets). One µL of first-stage product was used as template without cleanup, with 2 µL of primer per reaction. Cycling conditions were 98°C for two minutes, followed by eight cycles of 98°C for 10 seconds, 60°C for one second, and 68°C for 1 second. PCR products were pooled in equal volumes, purified using a 0.6X AMPure ratio, and sequenced with a 10% PhiX spike-in on an Illumina MiSeq i100+ instrument (Illumina, San Diego, CA, USA) using a 600-cycle kit with read lengths extended to 2 x 309 bases.

### 16S Error Rate Analysis

Reads were mapped against the reference 16S and 18S rRNA gene sequences provided for the ZymoBIOMICS Microbial Community DNA Standard, after removing duplicate gene copies and masking positions at which distinct copies of the same species differed; masked positions were excluded from both the numerator and denominator of all error rate

calculations. Paired-end reads were merged using BB-Merge,<sup>14</sup> exploiting the overlap of the V4–V5 insert to suppress independent sequencer errors by consensus; read pairs without a confident overlap were discarded. Merged reads were trimmed to remove primer sequences using Cutadapt<sup>15</sup> with the `–discard-untrimmed` flag, and reads shorter than 100 bp after trimming were discarded. Trimmed reads were aligned to the deduplicated reference using BWA-MEM<sup>10</sup> and coordinate-sorted using SAMtools<sup>12</sup>; unmapped, secondary, and supplementary alignments were excluded. Per-base error counts were extracted from sorted, indexed BAM files using a custom Python script (pysam<sup>12</sup>). Substitution events were classified in the canonical pyrimidine context by strand collapse, enumerating six substitution classes (C>A, C>G, C>T, T>A, T>C, T>G); C-context rates were computed as events divided by total covered C and G positions (excluding masked positions), and T-context rates analogously over T and A positions. Insertion and deletion rates were computed as the number of inserted or deleted bases divided by the total number of covered reference positions, excluding masked positions.

## Mutational Signature Analysis

Variant positions, reference and derived alleles from condition-specific VCF files were processed using the Signal mutational spectra elicitation webservice for single and double substitutions and indels (<https://signal.mutationalsignatures.com/>).<sup>16,17</sup> Signature reconstruction was first performed on the absolute mutational spectra of each condition using SigProfilerAssignment (COSMIC v3.4, default activity thresholds).<sup>18</sup> Dye-imparted mutational spectra were then assessed by subtracting the frequency of each variant class in the no-dye control from the corresponding frequencies in each of the three dye-treated samples. These difference frequencies were multiplied by the number of observed sequence variants of the corresponding class in each dye-treated sample, rounded, and floored to zero, then submitted as 96-, 78-, or 89-channel variation catalogs to Signal. Signature reconstruction was repeated on these different spectra using SigProfilerAssignment with the same parameters. Indel classes were assessed using a recently published taxonomy.<sup>19</sup> For base count level processing, non-reference sites were annotated using the Signal's companion library signature.tools.lib (<https://github.com/Nik-Zainal-Group/signature.tools.lib>), including the indel.sig.tools.lib-contained functions in the R statistical environment (v. 4.2.0). Corresponding reference and derived allele counts were retrieved from the original VCF files and tallied as the above variation catalogs for downstream processing in Signal.

## Results

To evaluate the gross impact of intercalating dyes present in the amplification reaction on NGS library quality, P5/P7-anchored libraries were prepared from enzymatically sheared human genomic DNA (GIAB reference sample

NA12878) in the presence of SYBR Green I (Thermo Fisher Scientific), EvaGreen (Biotium), or EvaGreen Plus (Biotium) using NEB Q5 High-Fidelity DNA Polymerase on an iconPCR instrument (n6tec) (Figure 1A). Relatively high concentrations of dye and cycle counts that exceed the standard stop point of the iconPCR instrument (n6tec) were selected to amplify any mutagenicity from the dyes while remaining within its performance envelope. Libraries were sequenced on an Element AVITI (Element Biosciences) to maximize sensitivity to any dye-induced mutations. Global library metrics showed that dye-containing reactions produced libraries of comparable quality to the no-dye control, with only subtle differences in insert size distribution (Figure 1B) and GC content (Figure 1C) observed.

The primary impact of intercalating dyes was expected to be in the induction of mutations. Single-base mutational rates were characterized relative to the total number of bases sequenced (since all errors are expected to be technical in origin). Mutation analysis found very similar error rates among all the samples, regardless of the presence of dyes, with dye-containing samples increasing the error rate by less than 3.5% above the no-dye control, with SYBR Green I (Thermo Fisher Scientific) showing slightly lower substitution error rates relative to EvaGreen (Biotium) formulations. Base-by-base substitution errors showed similar rates of transitions and transversions (Figure 1D). Indel mutation rates were also only slightly perturbed by the presence of the dyes with ~5% more deletions and 1–2% additional insertions observed, but still extremely close to background rates (Figure 1E).

Because many sequencing applications require higher cycle counts than the previously mentioned genomic libraries, we next explored whether dye-induced errors might become more pronounced under extended amplification. An *rRNA* gene amplicon control (ZymoBIOMICS Microbial Community DNA Standard) was amplified for 16, 20, 24, or 28 cycles in the presence or absence of SYBR Green I (Thermo Fisher Scientific), and single-base and indel error rates were quantified (Figure 2A). Observed error rates rose markedly with increasing cycle number across all substitution and indel classes, with the highest rates seen at 28 cycles (Figure 2B,C). Against this backdrop, the contribution of SYBR Green I (Thermo Fisher Scientific) was detectable but small, with dye-containing reactions at matched cycle counts showing only slightly elevated error rates relative to the no-dye control. The reduction in cycle-driven error afforded by the reduced cycle counts of the iconPCR instrument (n6tec) should therefore more than offset the small additional error associated with the presence of dye.

We next investigated changes in mutational signatures caused by the presence of dyes in the amplification reaction. Mutational signatures are more sensitive than SNP analysis in that they include the context of the nucleotide change and are frequently highly correlated with the mechanism of action.<sup>7</sup> Thus, a mutation that is associated with base intercalation would be expected to have a specific signature which could help separate it from the background. Since all observed variants are expected to be technical errors in this reference sample design, any signature enriched

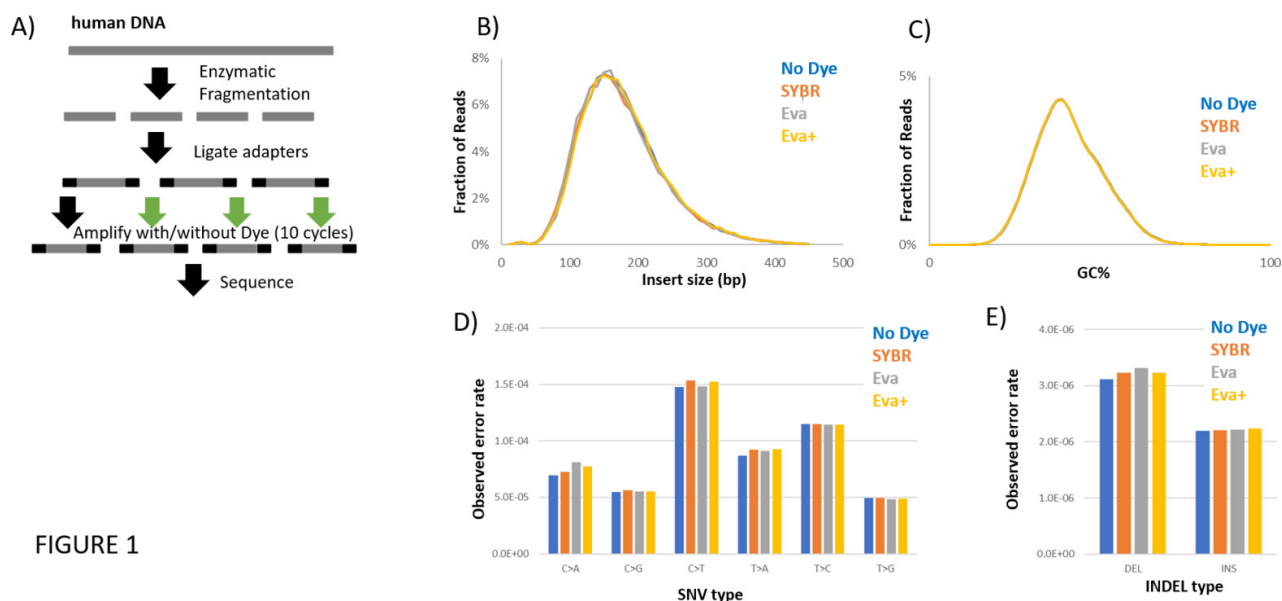


FIGURE 1

**Figure 1. Impact of intercalating dyes on NGS library quality and mutation rates.** (A) Schematic of the experimental workflow. (B) Histogram of insert lengths (bp) for libraries prepared under each dye condition. SYBR: SYBR Green I; Eva: EvaGreen; Eva+: EvaGreen Plus. (C) Frequency distribution of GC fraction for libraries prepared under each dye condition. (D) Observed single-base mutational error rates by substitution type (C>A, C>G, C>T, T>A, T>C, T>G), normalized to total bases sequenced, for libraries prepared under each dye condition. (E) Observed indel error rates by event type (deletion, insertion), normalized to total bases sequenced, for libraries prepared under each dye condition.

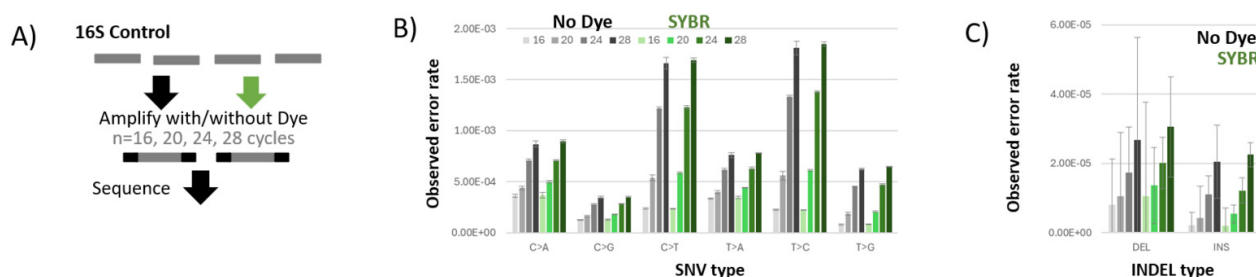


FIGURE 2

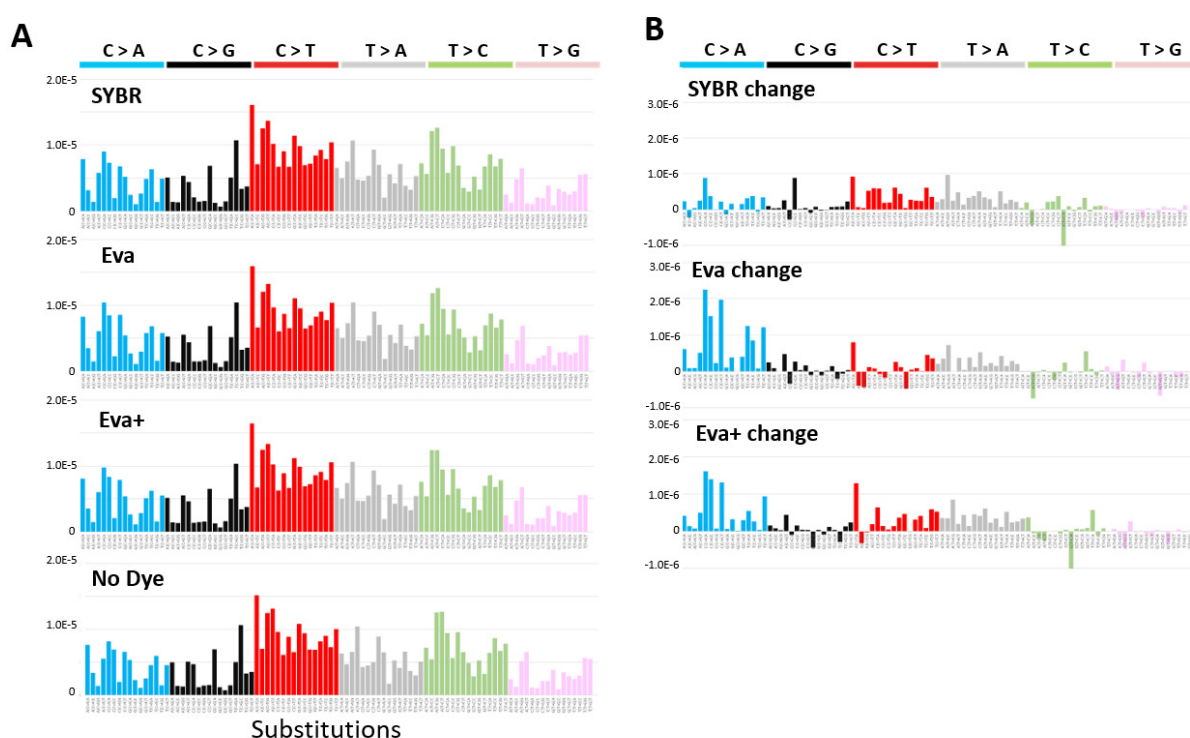
**Figure 2. Impact of cycle number and intercalating dye on mutation rates in 16S amplicon sequencing.** (A) Schematic of the experimental workflow. (B) Observed single-base mutational error rates by substitution type (C>A, C>G, C>T, T>A, T>C, T>G), normalized to total bases sequenced, for each cycle number (16, 20, 24, 28) with and without SYBR Green I. No dye: gray; SYBR Green I: green. (C) Observed indel error rates by event type (deletion, insertion), normalized to total bases sequenced, for each condition as in (B).

in dye-containing conditions would be attributable to the dyes themselves.

Mutational spectra across all dye conditions were superficially indistinguishable from the no-dye control (Figure 3A). Attempted decomposition of these spectra into known SBS (single-base substitution) signatures using SigProfilerAssignment revealed no dominant signature attributable to dye exposure; since this reaction system is entirely in vitro, a genuine dye-induced mutagen would be expected to produce a simple and recognizable signature. Reconstruction from known SBS types confirmed no systematic difference between conditions. To increase sensitivity, the no-dye spectrum was subtracted from each dye-containing

condition to isolate any residual dye-specific signal (Figure 3B). These difference spectra showed only modest increases, predominantly in C>T and T>C transitions across all dyes, with EvaGreen (Biotium) and EvaGreen Plus (Biotium) also showing minor C>A transversions. Attempted signature reconstruction of the difference spectra identified no dominant signal; the highest contributors were SBS4 (tobacco smoke-associated) in EvaGreen (Biotium) (0.54) and SBS8 (unknown etiology) in SYBR (Thermo Fisher Scientific) and EvaGreen Plus (Biotium) (0.486 and 0.421, respectively), with no consistent pattern across conditions. Thus, the differential nucleotide transition signature represents a likely novel artifactual signature.

FIGURE 3



**Figure 3. Single-base substitution spectra and dye-induced changes in libraries amplified with intercalating dyes.** (A) Observed mutational spectra across 96 trinucleotide substitution contexts, normalized to total bases sequenced, for libraries prepared under each dye condition and the no-dye control. Substitution classes (C>A, C>G, C>T, T>A, T>C, T>G) are color-coded as indicated. SYBR: SYBR Green I; Eva: EvaGreen; Eva+: EvaGreen Plus. (B) Dye-imparted substitution changes, calculated by subtracting the no-dye spectrum from each dye-containing condition.

We next looked for changes in indel signatures associated with dye presence in the amplification reaction (Figure 4A). Given that an intercalated dye molecule occupies a position structurally analogous to a base pair, intercalating dyes might be expected to preferentially induce single-base insertions, particularly in repetitive sequence contexts where slippage is most likely. Absolute indel spectra were highly similar across all conditions with no major shift in pattern observed. Consistent with the bulk indel rates observed in Figure 1, deletion events exceeded insertions in all dye conditions; notably, no enrichment of single-base insertion events was observed in any specific sequence context, ruling out a context-specific insertion signal masked by the overall deletion excess. Focusing on the differences between the spectra (Figure 4B) revealed modest but consistent shifts in indel patterns across dye-containing conditions. The most prominent signals were a modest increase in 2–4 bp deletions (~4–10% above the no-dye control, with EvaGreen (Biotium) being the most pronounced) and deletions of a single C preceding an A (~4–10% above the no-dye control). A corresponding decrease in deletions of 5 bp or greater (~2–6% below the no-dye control) was also observed. In all cases, the changes represented a minority of observed mutations.

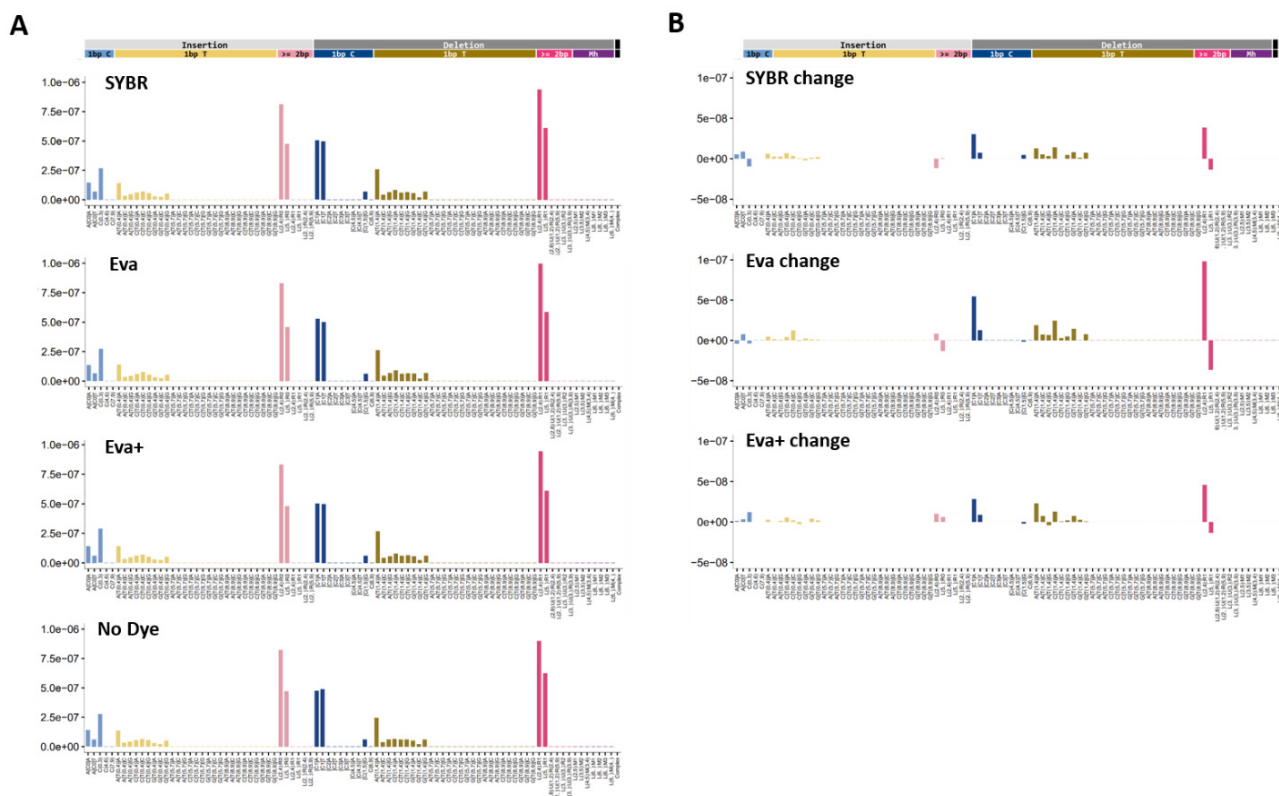
We next examined whether the observed indel changes corresponded to any known mutational signatures. Recon-

struction of the absolute indel spectra identified no dominant signature attributable to dye exposure, with the sum of all assigned signatures reaching a cosine fit of only 0.77 across conditions. An analysis of the difference spectra (Figure 4B), where only dye-induced changes should be represented, yielded even poorer overall fits (0.67–0.71). The largest contributors to these difference spectra included signatures associated with TOP1-linked deletions (Ind4a) and APOBEC activity (Ind9a), neither of which has a plausible mechanistic link to base intercalation. Attempted modeling against a panel of experimental chemical genotoxin signatures similarly failed to identify a coherent signal, with the highest-scoring match, 6-Nitrochrysene, explaining only a minority of the EvaGreen (Biotium) difference signatures, and cosine fits ranging from 0.45–0.56 across conditions. The indel difference spectrum therefore lacks any coherent signature attributable to dye exposure, consistent with the substitution analysis.

## Discussion

Intercalating dyes present during iconPCR amplification had minimal impact on mutational rates across all analyses performed. In high-fidelity genomic library preparation, single-nucleotide variant and indel rates in dye-containing conditions were close to the background error rate of the

FIGURE 4



**Figure 4. Indel spectra and dye-induced changes in libraries amplified with intercalating dyes.** (A) Observed indel spectra across insertion and deletion contexts as defined by Koh et al.,<sup>19</sup> normalized to total bases sequenced, for libraries prepared under each dye condition and the no-dye control. Complex mutations are not shown. SYBR: SYBR Green I; Eva: EvaGreen; Eva+: EvaGreen Plus. (B) Dye-imparted indel changes, calculated by subtracting the no-dye spectrum from each dye-containing condition.

no-dye control, with only modest increases in short deletions and no enrichment of single-base insertions, and mutational signature analysis revealed no coherent signature attributable to dye exposure in either substitution or indel space. Extending the analysis to higher cycle counts in a 16S amplicon system reinforced this conclusion: while observed error rates rose substantially with cycle number, the additional error attributable to the dye remained small at every cycle count tested. Together, these experiments indicate that cycle number, rather than the presence of dye, is the dominant driver of accumulated amplification error. Because these results were obtained under deliberately stringent conditions, with dye concentrations and cycle counts set above the standard iconPCR operating parameters, the rates observed here likely represent an upper bound on real-world dye-induced mutagenicity. One practical advantage of the iconPCR instrument (n6tec) is therefore its reduction in the number of cycles required for many applications, which itself mitigates the formation of artifactual mutations and offsets the very low levels of dye-derived error observed here.

The observed error rates are unlikely to meaningfully impact results in the large majority of NGS workflows, including bulk whole-genome and whole-exome sequencing, RNA-seq, ChIP-seq, and ATAC-seq. Applications that de-

pend on variant detection at the lowest measurable allele frequencies warrant more careful consideration. Minimal residual disease detection requiring discrimination of somatic variants at allele frequencies as low as 0.001–0.01%,<sup>20</sup> where any systematic elevation in technical error rates could contribute false-positive calls, is one such domain. Single-cell DNA sequencing for clonal analysis, where amplification errors cannot be averaged across a bulk population and are a recognized source of noise,<sup>21</sup> represents another such application. Finally, de novo mutational signature analysis in translational or mechanistic contexts warrants particular awareness. The consistent pattern of C>T and T>C transitions in the substitution difference spectra, together with the modest enrichment of 2–4 bp deletions in the indel difference spectra, provides a recognizable fingerprint that should make any dye contribution detectable at the level of the mutational spectrum, though it cannot address the origin of any individual variant. Accordingly, caution is warranted in the most sensitive applications operating at the limits of detection, where even a modest shift in the error background may have meaningful consequences for the results.

## Acknowledgments

The authors are thankful for Drs. John Essigmann, Robert Croy, Bogdan Fedeles, and members of the MIT BioMicro Center for sequencing and their constructive comments and discussions on this manuscript. We acknowledge the efforts of Cecilia Chau from the Genomics and Microbiome Core Facility (GMCF; RRID:SCR\_026856) at Rush University Medical Center for performing library preparation and sequencing of *16S rRNA* gene amplicons.

## Financial Support / Conflict of Interest

This work was funded by the National Cancer Institute of the US National Institutes of Health (NIH) under award P30-CA14051. Data from the manuscript are available at <https://fairdomhub.org/studies/1454> including links to all data repositories. Pranav Patel is an employee of n6 bioTec.

Submitted: March 25, 2026 EDT. Accepted: June 05, 2026 EDT.

Published: September 01, 2026 EDT.



This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CCBY-4.0). View this license's legal deed at <http://creativecommons.org/licenses/by/4.0> and legal code at <http://creativecommons.org/licenses/by/4.0/legalcode> for more information.

## References

1. LePecq JB, Paoletti C. A fluorescent complex between ethidium bromide and nucleic acids: physical-chemical characterization. *J Mol Biol.* 1967;27(1):87-106. doi:[10.1016/0022-2836\(67\)90353-1](https://doi.org/10.1016/0022-2836(67)90353-1)
2. McCann J, Choi E, Yamasaki E, Ames BN. Detection of carcinogens as mutagens in the Salmonella/microsome test: assay of 300 chemicals. *Proc Natl Acad Sci U S A.* 1975;72(12):5135-5139. doi:[10.1073/pnas.72.12.5135](https://doi.org/10.1073/pnas.72.12.5135)
3. Singer VL, Lawlor TE, Yue S. Comparison of SYBR Green I nucleic acid gel stain mutagenicity and ethidium bromide mutagenicity in the Salmonella/mammalian microsome reverse mutation assay (Ames test). *Mutat Res.* 1999;439(1):37-47. doi:[10.1016/s1383-5718\(98\)00172-7](https://doi.org/10.1016/s1383-5718(98)00172-7)
4. Aird D, Ross MG, Chen WS, et al. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.* 2011;12(2):R18. doi:[10.1186/gb-2011-12-2-r18](https://doi.org/10.1186/gb-2011-12-2-r18)
5. Kebschull JM, Zador AM. Sources of PCR-induced distortions in high-throughput sequencing data sets. *Nucleic Acids Res.* 2015;43(21):e143. doi:[10.1093/nar/gkv717](https://doi.org/10.1093/nar/gkv717)
6. Bell DA, DeMarini DM. Excessive cycling converts PCR products to random-length higher molecular weight fragments. *Nucleic Acids Res.* 1991;19(18):5079. doi:[10.1093/nar/19.18.5079](https://doi.org/10.1093/nar/19.18.5079)
7. Alexandrov LB, Nik-Zainal S, Wedge DC, et al. Signatures of mutational processes in human cancer. *Nature.* 2013;500(7463):415-421. doi:[10.1038/nature12477](https://doi.org/10.1038/nature12477)
8. Zook JM, Catoe D, McDaniel J, et al. Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci Data.* 2016;3:160025. doi:[10.1038/sdata.2016.25](https://doi.org/10.1038/sdata.2016.25)
9. Zook JM, McDaniel J, Olson ND, et al. An open resource for accurately benchmarking small variant and reference calls. *Nat Biotechnol.* 2019;37(5):561-566. doi:[10.1038/s41587-019-0074-6](https://doi.org/10.1038/s41587-019-0074-6)
10. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv.* 2013;1303.3997v2. doi:[10.48550/arXiv.1303.3997](https://doi.org/10.48550/arXiv.1303.3997)
11. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics.* 2011;27(21):2987-2993. doi:[10.1093/bioinformatics/btr509](https://doi.org/10.1093/bioinformatics/btr509)
12. Danecek P, Bonfield JK, Liddle J, et al. Twelve years of SAMtools and BCFtools. *GigaScience.* 2021;10(2):giab008. doi:[10.1093/gigascience/giab008](https://doi.org/10.1093/gigascience/giab008)
13. Naqib A, Poggi S, Wang W, Hyde M, Kunstman K, Green SJ. Making and sequencing heavily multiplexed, high-throughput 16S ribosomal RNA gene amplicon libraries using a flexible, two-stage PCR protocol. In: *Methods Mol Biol.* Vol 1783. Humana Press; 2018:149-169. doi:[10.1007/978-1-4939-7834-2\\_7](https://doi.org/10.1007/978-1-4939-7834-2_7)
14. Bushnell B, Rood J, Singer E. BBMerge—accurate paired shotgun read merging via overlap. *PLOS One.* 2017;12(10):e0185056. doi:[10.1371/journal.pone.0185056](https://doi.org/10.1371/journal.pone.0185056)
15. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* 2011;17(1):10-12. doi:[10.14806/ej.17.1.200](https://doi.org/10.14806/ej.17.1.200)
16. Degasperi A, Amarante TD, Czarnecki J, et al. A practical framework and online tool for mutational signature analyses show inter-tissue variation and driver dependencies. *Nat Cancer.* 2020;1(2):249-263. doi:[10.1038/s43018-020-0027-5](https://doi.org/10.1038/s43018-020-0027-5)
17. Degasperi A, Zou X, Amarante TD, et al. Substitution mutational signatures in whole-genome-sequenced cancers in the UK population. *Science.* 2022;376(6591):science.abl9283. doi:[10.1126/science.abl9283](https://doi.org/10.1126/science.abl9283)
18. Díaz-Gay M, Vangara R, Barnes M, et al. Assigning mutational signatures to individual samples and individual somatic mutations with SigProfilerAssignment. *Bioinformatics.* 2023;39(12):btad756. doi:[10.1093/bioinformatics/btad756](https://doi.org/10.1093/bioinformatics/btad756)
19. Koh GCC, Nanda AS, Rinaldi G, et al. A redefined InDel taxonomy provides insights into mutational signatures. *Nat Genet.* 2025;57(5):1132-1141. doi:[10.1038/s41588-025-02152-y](https://doi.org/10.1038/s41588-025-02152-y)
20. Newman AM, Lovejoy AF, Klass DM, et al. Integrated digital error suppression for improved detection of circulating tumor DNA. *Nat Biotechnol.* 2016;34(5):547-555. doi:[10.1038/nbt.3520](https://doi.org/10.1038/nbt.3520)

21. Valecha MK, Posada D. Somatic variant calling from single-cell DNA sequencing data. *Comput Struct Biotechnol J*. 2022;20:2978-2985. doi:[10.1016/j.csbj.2022.06.013](https://doi.org/10.1016/j.csbj.2022.06.013)